UNIVERSITY OF JYVÄSKYLÄ

WORD AND ITEM FEATURES THAT AFFECT MEASUREMENT OF VOCABULARY SIZE IN A MULTIPLE CHOICE VOCABULARY TASK

A Pro Gradu Thesis

by

Timo Törmäkangas

Department of English
1999

HUMANISTINEN TIEDEKUNTA
ENGLANNIN KIELEN LAITOS

Timo Törmäkangas

WORD AND ITEM FEATURES THAT AFFECT~~S~~ MEASUREMENT OF VOCABULARY SIZE IN A MULTIPLE CHOICE TASK

Pro gradu-tutkielma
Englantilainen filologia
Heinäkuu 1999 105 sivua + 10 liitettä

Sanaston kokoon vaikuttavat monet tekijät, minkä vuoksi testaaminen saattaa vaikeutua. Tutkimuksissa on havaittu useiden tekijöiden (esim. sanan yleisyys, sanan pituus ja sanaluokka) vaikuttavan sanojen vaikeuteen, ja siten siihen miten sanat opitaan. Jotta sanaston koon arviointi voitaisi laatia tehokkaksi ja oikeudenmukaiseksi, olisi testaajien oltava perillä sanaston vaikeuteen vaikuttavista tekijöistä.

Tutkielman tarkoituksena oli selvittää millaisin piirtein sanaston laajuutta mittaavaa monivalintatehtävää voidaan parhaiten luonnehtia. Aineistona käytettiin sanaston kokoa mittaavaa tehtävää, jota käytettiin tehtävänä Yleisissä kielitutkinnoissa vuonna 1996. Selvitettävät piirteet liittyivät tehtävien sisältöön ja toisaalta taas empiirisiin tunnuslukuihin, jotka voitiin laskea tehtävä suoritusten perusteella. Koska aineisto oli laaja, ja tarkoituksena oli tehdä yleistäviä päätelmiä tehtävän perusteella, tutkimus on kvantitatiivinen.

Tarkasteltavat piirteet määräytyivät pääasiassa testissä käytettyjen sanojen ominaisuuksista ja itse osioiden piirteistä. Sanojen ominaisuuksia voidaan käyttää hyväksi tehtävän laadinnassa kun pyritään ennakoimaan testin vaikeustasoa. Empiiriset osioiden piirteet taas kertovat testin havaitusta vaikeustasosta, sekä arvioinnin luotettavuudesta.

Pääkysymyksen ohella arvioitiin myös testin laatu ja sen rakenne sekä mittaamien ulotteiden määrä. Testi havaittiin jokseenkin yksipuoliseksi, koska tehtävät koostuivat suureksi osaksi substantiiveista. Jotkin muuttujat (semanttinen luokitus) osoittivat puolestaan, että testi oli niiden suhteen liian monipuolinen, jotta näitä muuttujia olisi testin kohdalla analysoida. Testiä voitiin pitää kolmen eri perusteen johdosta yhtä ulottuvuutta mittaavana, jonka johdosta testiä voidaan pitää validina.

Analyysin tuloksena voitiin havaita, että osioiden sisältöön pohjautuvien muuttujien perusteella tehtävissä menestyminen pystyttiin ennustamaan täydellisesti. Empiirisiin piirteisiin perustavat muuttujat pystyivät ennustamaan menestymisen johdonmukaisemmin viiden suurimman varianssilähteen ollessa kyseessä. Johtopäätöksenä voidaankin sanoa, että testin laadinnassa sisältöön perustuviin piirteisiin tulisi kiinnittää huomiota, jotta testi voidaan laatia johdonmukaisesti ja validiksi mittariksi. Tässä tutkielmassa käytetyt piirteet muodostivat hyvän ryhmän, koska niiden perusteella ennustus tehtävissä suoriutumisesta onnistui jokaisessa tapauksista.

Asiasanat: test theory, multi-dimensional scaling, discriminant analysis, vocabulary size, measurement, quantitative analysis.

CONTENTS

1 INTRODUCTION

Educational measurement is concerned with operationalised conceptions of proficiency. Proficiency scaling is a way to describe progressive degrees of ability in terms of a single aspect of a skill and to relate it through empirical research to numerical representation. The empirical link between a quantity and ability can be established via observation in the form of research into classroom activities, language abilities or other such activity. The objective is to describe proficiency in terms that can be observed and measured. Proficiency scales are important for several purposes, as they are a way of indicating the learners' relative mastery of a subject area, and thus enable efficient placement of persons on courses or work tasks.

The assessment of language abilities requires careful test construction in order to estimate abilities in question. Measurement that relates to real-life activities rather than a specific curriculum is sometimes referred to as criterion-referenced measurement (e.g. Hambleton 1988). A number of problems relating to criterion-referenced measurement include specification of a useful test content in terms of a scale of proficiency. While on the face of things, it seems easier to construct a general scale of proficiency for the communicative skills (e.g. reading), it seems more difficult to find defining features of a scale for a skill such as vocabulary knowledge which would seem to be rather idiosyncratic in terms of acquisition. Hence a scale of vocabulary size does not have importance only for assessment purposes, but also for vocabulary teaching and learning purposes, in that at best it would enable the use of material suitable for different kinds of people in different proficiency levels.

The construction of such a scale, however, requires the use of a reliable measure with a representative sample of subjects. When this is the case, the results can be generalised. This study was conducted on a test of vocabulary size, which was administered as a part of Finnish National Certificate language examinations. The aim was to identify those features that characterise the performance on the vocabulary tasks, and which establish a useful link between the instrument and a measurement scale. This was done by discovering the structure observable in the set of vocabulary items and contrasting it with the structure of the performance data.

2. SPECIFICATION OF THE DOMAIN OF VOCABULARY SIZE

An important part of establishing a measure of something is to specify what that something is. For vocabulary size measurement this means examining the definition of a word and vocabulary, what is difficulty in acquiring a vocabulary, and how it can be adequately and accurately measured.

2.1 Definition of word

The word can be defined from different points of view. Definition of a word in its widest sense takes into account the form, meaning and use of the word, but another related activity is classification of words. For practical purposes these two may sometimes be used synonymously, because even though the word can be defined on an abstract level, it is not always useful to stick to that definition. So, while definition attempts to capture a whole domain, classification may mean limiting the domain to an accessible frame of reference.

The word is formally defined as a sequence of letters separated by spaces or punctuation marks (see e.g. Johansson and Hofland 1989: 7 and Carter 1987: 4). Thus the word has a clear orthographic definition in the written language. Crystal (1987: 91) indicates that the formal definition may also be linked to words when they are spoken. Most speakers tend to put pauses between words rather than within words. He also suggests that, if words are added to a sentence, they are in most cases added between words instead of within them (however, see Crystal 1987: 91 for exceptions).

It is also important to recognise that while some lexical units meet the formal criterion of word, they are not necessarily classified as words. Goulden, Nation and Read (1990: 343) would exclude abbreviations, names of persons and places, prefixes and suffixes from the group of words. These words may be linked to context or culture-specific ideas, and are therefore well suited for assessment or other educational purposes. Also, Lyons (1968: 197) further defines lexeme in a way that may help to include different forms of words. Lexeme is a abstract term which refers to all of the inflected forms of a word. This definition is useful in that it subsumes the different inflected forms of a word, and it also covers expressions that are formed of several words (Jackson 1991: 11). Here the emphasis also turns to what the words mean. Arabski (1998: 26) notes that proper names are more difficult to retain, because they do not have the same meaningful properties as other lexical items, and for the beginner e.g. *baker* as a profession is easier to learn that the name *Baker* (Cohen & Burke 1993:250), because on a general level the profession can be related to concrete activities, while it is not the case with the proper name.

The formal definition is not necessarily adequate from the communicative point of view. Words are not only pieces put one after another to form sentences, but they are connected in a systematic way to convey information to others. Words are used to refer to abstract or concrete objects such as activities, emotions or tangible things, and in this way they mean something. This semantic function of words is to establish a reference between a sequence of letters and something either real or imaginary in the world. Hurford and Heasley (1990: 25) say that reference is a way for a speaker to indicate which things in the world are being talked about. Thus, for instance the lexical item *book* refers to a collection of pages that is intended for reading. Crystal (1987 : 91) also cites Bloomfield who sees words as the smallest units of speech that can meaningfully stand on their own, but also criticises this against the fact that, according to this view, for instance articles would be counted as words. Hammerly (1982: 452) would make a distinction in vocabulary based on the reference relationship of words. He would include content words (nouns, verbs, adjectives and adverbs), phrases and idioms into a vocabulary of a language, but he would exclude function words (pronouns, auxiliary verbs, determiners, prepositions, conjugations etc.) as part of grammar, because they express grammatical meaning, and relate more to the domain of grammatical knowledge rather than word knowledge. Words thus differ in which things in the word are being referred to.

Nation (1990: 53) calls the object of reference a concept. Concept is something that is formed by all the specific objects that can be referred to with a certain word. For example, the concept of *ball* shares all the common features of specific balls a person has encountered, such as *roundness*, *bouncing motion*, etc. Bourne et al. (1986: 126) have examined this from the point of view of cognitive psychology and claim that it is the concepts or representations that are stored in memory rather than the actual objects encountered. Subsequently, what is carried in memory is not the actual experiences or words, but their representations or categories, which seems to relate to what Nation refers to with 'meaning'. Nation (1990: 52) defines the relationship of a word and concept as that of translation. A speaker or writer therefore chooses to use a certain word to refer to an idea that he has, and uses that word, because it has the appropriate reference relation to transfer the idea into concrete words. Notable problems with the theory are that this translation is not exactly comparable across languages, and the translation even on the translation to mother tongue is indirect, so that it is difficult to assess it (Nation 1990:52).

In order to be more precise in the definition of the semantic system in a language, the distinction made between reference and sense has to addressed. Hurford and Heasley define *sense of an expression* as the place of an expression in the system of semantic relationships with the system of other expressions in the language (1990: 28). They say that every expression has a sense, but not necessarily a referent in the world, as is the case with for example connectives. Luria (1982: 36-51) also defines the word according to meaning. He sees the word as a unit that classifies human experience, and distinguishes *meaning* and *sense*. Meaning refers to the whole range of things a word can refer to, covering all the possible contexts in which the word can be correctly used. Sense then refers to a particular meaning of a word that is correct in a given context.

The semantic definition of the word can sometimes be in disagreement with the formal definition. For instance, phrasal verbs present such a case, where it is difficult to say if two (or more) words constitute a single concept or if several distinct meanings are in question. In practice expressions such as *mix up* can be translated with one word in another language (or vice versa) depending on the language's tendency to express things in the world, and thus it may seem that the expression contains a single concept. One way to deal with expressions such as this is to not count them as words, and exclude them into a domain of their own. This, however, will complicate the definition of what it is to know language.

It is also important to note that the words themselves are not the smallest meaningful units in a language. Words can be formed of parts which themselves have a meaning. Some of these morphological units refer more clearly to a kind of meaning an actual word may have. For example, the word *careless* is characterised by the suffix *-less* which indicates lack of something. While it tells something about the meaning of the word it does not indicate the exact meaning of the word. Goulden et al. (1990: 345) would exclude derived and inflected word forms from the count (as opposed to definition) of words, and they suggest using Nagy and Anderson's definition of levels of derivation as a criterion in distinguishing derived word forms. This word classification defines five levels of derivation based on the semantic relatedness between two words (Nagy and Anderson 1984: 309), and in it those words that are not derived or inflected are called base words. This distinction has been used in several studies of vocabulary size (see e.g. Nation 1993).

2.2 Definition of vocabulary

Vocabulary is defined as a collection of words. A vocabulary of English language, for instance, consists of a countless number of words with new ones added continuously. A vocabulary has two important dimensions, size and relationships between the words in it (Meara 1996). Usually measures of vocabulary size have been of central focus in criterion-referenced measurement, as the number of words known has been noted to be a good indicator of language proficiency in general correlating with other skills, e.g. reading comprehension (see e.g. Thorndike 1973). It is of course natural to think that with a large vocabulary communication is easy, and more difficult with a small vocabulary. The structure, or the relationships between the words in a vocabulary, is also important, and has a potential effect on the size of a learner vocabulary. Words exist in semantic relationships to each other. The relationships are hierarchical and words can exist in horizontal relationships of inclusion (hyponymy) or vertical relationships (synonymy, antonymy).

For practical reasons, teaching and testing can use a definition of vocabulary that is restricted in some manner, and focuses on the most important or useful words only. One way to do this is to define word according to the aspects discussed in the previous chapter, but even if the word count excludes certain words on the basis of formal grounds, the number of words in the vocabulary may still be quite high. For this reason, it is more useful to classify which are the important aspects that should be taken into account, when defining a vocabulary. This means that the definition may have to take into account the interests and / or needs of the target group. If the learners or examinees are students of engineering, they will most likely need to know technical vocabulary in addition to words needed for survival in everyday life. However, if the target group comes from different walks of life, the words used should be such that they do not contain entries requiring specialist knowledge. Similarly learners on different levels of proficiency need to know different kinds of words, because they are expected to manage with different degrees of success in different situations.

Vocabulary definition may also take into account the proficiency level of the learners. For this reason, there have been attempts to define difficulty levels of words or a basic set of words that would enable efficient vocabulary instruction. Another approach is to define a closed set of words through semantic fields. Lyons (1977: 253) defines semantic fields as consisting of words related to some conceptual field. In other words, words in a semantic field are more closely related to one particular, conceptually similar area of life than to others.

Nation (1993) suggests that for assessment general purposes, an adequate vocabulary can be defined either with the help of a dictionary and a frequency count. Words are first selected from a dictionary, and they can then be used for teaching or testing in an appropriate way taking into account their suitability for the learners. This definition helps to define a narrower vocabulary for assessment purposes by supplying a direct frame of reference from which actual words can be selected for testing purposes.

2.3 Word knowledge and acquisition

Vocabulary acquisition is a complex process, because the learner has to master many aspects of knowledge related to a word in order to communicate effectively (native-like performance is sometime used as a yardstick). Knowledge of a word is composed of several aspects related to a word. Nation's (1990: 33) is probably the best known description of aspects of knowledge related to a word. These are shown in table 1.1. A division made often is to distinguish receptive and productive knowledge. Receptive knowledge relates to knowledge about aspects of the word, when the word is heard or read, and productive knowledge relates to the productive skills speaking and writing.

Form is divided into spoken and written. This refers to the medium in which a word is encountered. Position refers to the place of the word in a grammatical system and the kinds of words with which a word is used or in which it can be encountered. Nouns, for example, have a certain place in which they can appear in a sentence, and they also conjugate according to number. Knowledge of collocation refers to relationship in which a word co-occurs with another. Some words can have several collocations (e.g. sun shines, sun rises, etc.) while others are restricted to just a few. The functions of a word refer to the frequency of occurrence of the word in a language, and also the appropriateness of a word in a given context. From the communicative point of view the knowledge of meaning is the most important. It has already been discussed the that relationship of a word form and its meaning (or concept) is necessary for communicating ideas. A word's associations, i.e. the kinds of feelings and ideas related to a word, are also included under meaning. The distinction is made between breadth and depth of knowledge (Read 1993: 357) emphasising the number of words a person~~s~~ knows as opposed to how well he knows them (i.e. the extent to which a person is able to map different knowledge relationships for a word).

Form	Spoken form	Receptive
		Productive
	Written form	Receptive
		Productive
Position	Grammatical position	Receptive
		Productive
	Collocations	Receptive
		Productive
Function	Frequency	Receptive
		Productive
	Appropriateness	Receptive
		Productive
Meaning	Concept	Receptive
		Productive
	Associations	Receptive
		Productive

Table 1.1. Aspects of word knowledge (Nation 1990: 31).

Meaning has a special emphasis in word knowledge. Underwood (1969) studied how information or experience is stored into long term memory, and claims that it is words in particular that can be associated with the information or experiences. To this effect, associating a word form to a meaning is probably most important for beginners, while other aspects may become more important at later stages of acquisition. Also, receptive and productive knowledge of denotation are not the same thing. In study of vocabulary size of comprehensive school learners Takala (1984: 205) found that the receptive vocabulary of Finnish comprehensive school students was larger than the productive vocabulary, although the difference was not high. An explanation for this would seem to be in the recommendation for teaching made by Hammerly (1982: 451) to focus on productive vocabulary skills on the initial level of learning. Later on the learners may benefit from a receptive knowledge of vocabulary regarding words that they do not need to use every day.

Several studies have been made to distinguish similarities and differences in acquiring a vocabulary as a mother tongue or as a foreign language. This interest has been especially important for research on universal grammar (see e.g. White 1989). The differences in learning outcomes may be seen in the success to which L1 and L2 learners manage to acquire a language. White (1989: 175) lists mother tongue, fossilisation, effectiveness of instruction and age as factors contributing to the acquisition process. The processes seem to be similar in the sense that learners in both cases go through similar phases of acquisition, i.e. initially a relatively large amount of words are learned, while the phase of learning decreases the further it progresses.

De Villiers et al (1978: 122) describe L1 learning of meaning as starting from concrete proper names. First, a one-to-one relationship is acquired, for example to use words like *Mommy* and *Daddy*. From this point the learner progresses to common nouns which require the learner to distinguish between specific and general reference of words. The learner faces a similar hurdle when learning verbs and adjectives. Then learning progresses in degrees to more abstract words. Foreign language learners also progress through relatively similar acquisition processes, and seldom make errors that fall outside of a predictable patterns of learning. Among others, Read (1993: 358) and Ellis (1994: 109) report evidence to support this. They, however, note that the patterns seem to indicate some kind of influence of the mother tongue.

The knowledge of languages is something that seems to separate language learners: L1 learner does not have previous knowledge of a language system, while FL learners do, and, depending on their mother tongue, also have different kind of exposure to language systems. This would lead one to assume that language acquisition is different for foreign language learners. Broeder et al. (1993: 48) found that language learners from different language backgrounds tend to have different preferences in constructing compound words from two nouns. The data that they collected would seem to indicate that learners rely quite heavily on their mother tongue in forming compounds.

Acquisition is also not simply directed by knowledge factors alone. Motivation is important in all learning, but also in learning words. Learners are interested of the type of vocabulary they think will be useful in the future. Therefore the lexicon that learners have and will have can reflect the context of their daily lives, e.g. professional needs; words that an accountant finds essential for survival in his profession are probably trivial for an engineer (although not necessarily). Therefore these learners probably follow different paths in vocabulary acquisition, and their vocabularies end up looking different. However, both vocabularies will also include a host of words that they share (for instance grammatical and more frequently used words), and these words are likely to be used in several contexts and generally appear frequently in individual texts.

2.4 The size of vocabulary

It has already been noted that vocabulary size can be divided into receptive and productive domains. What is the size of vocabulary in the receptive and productive domains, and what is their relationship? Röman et al. (1970: 24) quote two studies of vocabulary size of children. Nation (1990: 12) and Takala (1984: 50) cite similar figures, and the figures have been collected in table 1.3. However, where Nation's figures refer to base words known, it is not clear what measures Röman et al used.

The size of the vocabulary of adult native speakers have also been estimated. Hammerly (1982: 458) estimates that while a first grader has a command of about 17,000 basic and 30,000 derived words, an adult knows about 200,000 words. Nagy and Anderson (1984) and Hammerly (1982: 458) suggest that L1 learners add about 1000 to 2000 words to their vocabulary yearly. A foreign language learner cannot reach such a sizeable vocabulary, but he can communicate efficiently with a smaller number of words.

Age in years	Smith	Grisby	Terman and Childs	Kilpatrick
1	3			235
2	272			
3	896	1507		405
4	1540			700
5	2072	2527		1528
5.5			1528	
6	2562	3054		
6.5			2500	2500
7.5			2600	
8.5			3960	
8.5				4480
9.5			5000	
9.6				6620
10.5			6000	
10.7				7020
11.5			6100	
11.7				7860
12.5			7700	
12.8				8700
13.0			8800	
13.9				10660
15.0				12000

TABLE 1.3. Vocabulary size in different ages according to different studies.

Röman et al (1970: 25) suggest that the relationship between receptive and productive vocabularies is about 3:1. Hammerly (1982: 452), however, believes this ratio is not stable at all stages of learning. He believes that the learners on initial stage will benefit from productive vocabulary instruction, while on later proficiency levels the receptive vocabulary becomes more dominant.

What then is the size of these vocabularies in terms of the things a person can do? Carter (1987: 22, 1988: 2) suggests that a vocabulary of about 800 words is enough for a person to communicate basic needs in life. The relationship of vocabulary and reading has been researched the most because of a simple way of quantification. Also, Anderson et al. (1982: 245) see that in reading vocabulary measures have importance for readability. Also, it is easier to estimate the gain in knowledge when the advances in reading and vocabulary are easier to compute. Nation (1995: 693) found that the ease of reading unabridged novels is considerable when the vocabulary is about 2,500 words, and recommends 5,000 words to be used as a subsequent goal post in ease of reading.

2.5 Vocabulary acquisition strategies

The importance of acquisition strategies of words is central to testing, because they provide clues to the reasons, why certain words are more difficult than others, but also help to clarify what kinds of tests are appropriate for examinees in different proficiency levels. In the initial stages of learning vocabulary learning relies much on memorisation. Different teaching and learning methods, such as e.g. the key word method, have been developed to aid in effective memorisation. Cohen (1990:26) lists nine techniques that are shown in table 1.2. The variety of acquisition strategies available have been noted to related to the success of acquisition. Lawson and Hogben (1996) note that high-proficiency Australian learners of Italian were also more able to use different strategies to learn new words, and concluded that learning is not as much a matter of using a strategy to acquire words, but rather success depends more closely on selecting the right strategy for a given purpose.

1. By linking the word to the sound of a word in the native language, to the sound of a word in the language being learned, or to the sound of a word in another language.
2. By attending to the meaning of a part or several parts of the word.
3. By noticing the structure of part or all of the word.
4. By placing the word in the topic group to which it belongs.
5. By visualising the word in isolation or in written context.
6. By linking the word to the situation in which it appeared.
7. By creating a mental image of the word.
8. By associating some physical sensation to the word.
9. By associating the word to a keyword.

TABLE 1.2. Nine mnemonic techniques for remembering words (Source: Cohen 1990).

L2 learners are aware of one language, when they begin learning the second one, however, it is not clear to what extent the knowledge of having learnt one language helps or hinders learning the other one. For this reason, L2 learning usually starts from the assumption that it is easier to start acquisition from concrete words and then gradually move on to more abstract words as the learner becomes more comfortable with the new language. For concrete words, a direct reference (a picture or an object) can be used to illustrate the meaning of a word. Once the learner advances, it is possible to use textual means and inference from context. Also, as the size of vocabulary grows and the learner becomes aware of the structuring principles of the target language vocabulary, it is possible to use these rules and morphological rules to extend meanings to new words.

Use of context clues and morphological rules as a learning facilitator is a strategy used for more advanced students. Context can be used in two ways to aid vocabulary learning. Either the learner has no access to dictionary or other reference to help specify the meaning of a word, or he is allowed to use one. In the first instance, learners are encouraged to make a guess of the word meaning based on different clues they can pick out from the context where the word is met, e.g. a text. This may lead to vague knowledge of the word: the precise meaning of the word is not known, but knowledge of the context where the word can be expected to be encountered may offer partial clues to the meaning of the word. Lawson and Hogben (1996: 105) note that it is not necessary for a reader to fully comprehend the words in a text for it to be meaningful for him. In a study involving Australian students learning Italian, they found that when learners made a deliberate attempt to learn new words, they were most successful in acquiring word meaning from context, when they could use individual learning strategies rather than a fixed set of strategies (Lawson and Hogben: 1996, 127). However, Simonsen and Singer (1992: 210) note that there is evidence that context can also be used to aid in retention of words in reading comprehension tasks. They conclude that reading comprehension can be aided significantly by teaching ~~the~~ students word definitions.

Wysocki and Jenkins (1987: 66) studied school children's ability to derive word meanings using two sources of information. On the one hand they used morphological generalisation, and on the other they used context to aid in determining the meaning of a word. Their results showed that the ability was more pronounced for more able students, but that it did not improve the vocabulary scores of the learners, and also that they did not seem to be able to connect the information offered by the two sources.

In summary, thus, it would seem that individual learning strategies are favoured by current teaching of vocabulary. This means that learners will be equipped with fairly independent strategies in acquiring new words, which may ultimately show in the structure of their L2 lexicon and the kinds of words they ultimately acquire.

2.6 Difficulty in learning words

Difficulty has been defined for language teaching purposes in the following way: "When person A takes more time to learn item X than item Y, we may say that item X is more difficult for person A than item Y (Higa, 1965:168).

Viewed this way difficulty is not absolute, but relative in terms of the things being learned. This is an empirical way to describe difficulty, which may not be feasible in all contexts. Lado (1957) described a hierarchy of difficulty related to the teaching of words. His basis for the distinction of easy, normal and difficult words was the likeness of word forms in L1 and the target language. Nation (1990: 35) defines difficulty as *learning burden*, or the amount of effort that is needed to learn concept X. In terms of language ability, his conception of difficulty is equally related to the past and present experiences of the learner. Any aspect in the learning process that involves prolonging or increased effort for something new to be digested, also increases difficulty. This definition of difficulty would seem to be equally relativistic, but it provides a better basis for criterion-referenced measurement, because previous research can be used as a measure against which to measure development. In this view, for instance, Finnish learners of English can be viewed as a relatively homogenous group of learners as opposed to groups consisting of learners from different nationalities.

Several aspects affect the successful learning of a foreign language vocabulary. Some of them are objectively measurable, while others are not (e.g. personality, state of mind). Palmberg (1988: 207) divides the factors into intralingual, interlingual and extralingual factors. Features of words that have been researched include exposure (frequency of occurrence of the word in a language), distance of phonological and orthographic features between mother tongue and foreign language word, abstractness, word class, word length, and similarity of FL and native words (see e.g. Perkins 1987). Ellis and Beaton (1993: 568) also list familiarity of grapheme to phoneme mappings for spelling as a factor.

Nation's learning burden mentioned earlier is also one way of quantifying the difficulty of associated with the words (Nation 1990: 3). If a new word has no relation to what the learner has seen before, the learning burden is great. The more complex and foreign the word to be learned is, the more effort it requires to be learned. This is also a relational measure, and depends the learner's previous exposure to the language to be learned and other languages. It makes, however, easier to estimate~~d~~ the learning burden by investigating features of the words. In the following some of the word features and their link to difficulty of learning are described.

2.6.1 Distance to base words

Earlier it was mentioned that some word counts exclude inflected and derived word forms. Goulden et al (1990: 345) state that this division is important, because of the learning burden related to derived word forms in particular. The assumption is that it is not as difficult to learn derived forms of words as it is to learn completely new words. The affixes that a base word can have may also give a clue the learner about the meaning of a word. If the word, for example, has the suffix *-less*, and learner may from previous exposure to words ending with the same suffix know that the word has something to do with a lack of something.

Laufer (1989: 11), however, warns about deceptive transparency in connection of learning new words. Deceptive transparency applies to those words or expressions whose appearance does not directly lead to the meaning of the word. For example, *infallible* can be separated to constituents in + fall + ible, and a learner may be lead to think that the word means ”something that cannot fall” instead of the correct meaning. Deceptive transparency is more common with idioms and expressions, but applies also to words with deceptive morphological structure (Laufer 1989: 12).

2.6.2 Similarity of word forms

Palmberg’s interlingual feature refers to the effect of the first language on the acquisition process. Similarity of word forms refers to the status between words in two languages both in terms of form and meaning. The learning burden for word forms is considered low, if they are close to words that the learner already knows (Nation, 1990: 33). Word forms that are already familiar from the mother tongue are easy to include into one's vocabulary, while those that differ require more effort. The learning burden is increased when the differences are considerable, e.g. the writing conventions are different between languages.

Ellis and Beaton (1993: 559) studied learning paths of German learners of English in relation to the key word learning method, and found that the ease of learning is dependent on the similarity of the phonological and orthographic patterns of the native language. Their findings show that the closer the language systems of two languages are to each other, the easier it is for the learner to acquire the foreign language.

Ellis and Beaton examined the similarity of word forms by measuring the number of common letters and phonemes in a word in the two languages. This is however, a problematic measure, because some letters and phonemes may be rendered differently in the different communication modes. For example, the English word 'economic' may be translated as 'ekonominen' in Finnish. The sound patterns are fairly similar for the first part of the word /ekonomi/ and the notable difference can be observed at the end of the word. But the differences are not the same when the words are examined in the written or spoken formats.

The difference is important, because especially formal language learning takes place also in spoken interaction. It is also possible for a person to learn words through listening to spoken language, and then convert them fairly accurately into their written format. The orthographic pattern, however, is different in two places. It seems unclear which form of patterning should be used as the basis for classification.

It is also possible that a word in the foreign language resembles a word in mother tongue, but has a different meaning. While this may also be a source for confusion, it can also be seen as lowering the burden for integrating that new distinction into the vocabulary of the foreign language.

2.6.3 Word class

Different word classes are important for language, because they have different functions in a language, e.g. nouns are used as a label to refer to objects, and verbs refer to activity etc. It would be difficult to imagine performing any communicative activities without the knowledge of verbs, for instance. For this reason, it is also important that a vocabulary test contains an adequate sample of different word classes. It is also important to note that words have a proportionately different representation in a language. For example, Johansson et al. (1989: 15) estimate that in a corpus of over million words the word classes are represented as is shown in table 1.4.

Word class has importance for learning, and Ellis and Beaton (1993: 565) claim that it is easier to learn nouns and adjectives, while verbs and adverbs are more difficult. The reason for nouns being easier is usually attributed to similar aspects as concrete words, the referent can be relatively easily demonstrated.

Word class	Amount
Nouns	254,992
Verbs	179,975
Determiners	125,018
Prepositions	123,440
Adjectives	73,546
Conjunctions	71,498
Numerals	19,126

TABLE 1.4. Representation of word classes in LOB corpus.

Using nonsense words embedded in prose passages Na and Nation (1985: 33) found that among the factors that affect guessing of word meaning in context word class was an important factor. For the English teachers who participated in the study, verbs were easiest to learn, and following them were nouns, adverbs and adjectives.

2.6.4 Word length

The length of word has been noted to have an effect on the ease of learning a word. Long words are more difficult to learn, and this has been attributed e.g. to the strain the put on the memory (Arabski 1993: 24). Long words also tend to be less widely used in language. (Ellis and Beaton 1993:568).

2.6.5 Exposure

The exposure to words has been considered an important indicator of word difficulty (Nation 1990). The basic assumption is that if a learner is exposed to words, it is more likely that he will learn the meaning of those words. The more frequent the exposeure is, the stronger the link between the word and its meaning will become. This has been demonstrated in relation to reading. Readers can understand different amounts of texts depending on the size of their vocabularies. Nation (1992: 695) found that for pleasurable reading of most unsimplified texts a vocabulary of 2,000 most common words was not sufficient.

Nation proposes five levels of vocabulary size that relate to proficiency. The most basic level is the 2,000-word level, which enables reading of simplified novels. The 3,000-word level is where unsimplified texts can be read. The 5,000-word level marks a wide vocabulary, and 10,000-word level corresponds to a large wide vocabulary. The university word level is located between the two last levels, but corresponds to specialised vocabulary used in university texts. (Nation 1990: 263).

As a consequence, frequency lists of words have been thought to be useful for educational purposes. The representativeness of the frequency counts demand that the counts be based on the full range of variability in a language, which means that sources based on written sources only may misrepresent the word counts (Biber 1993: 243). Frequency counts also change over time. Some topics may dominate texts at one point in time, while other words are more frequent later. For this reason frequency counts should be based on as representative a sample as possible. The sources used for word selection should be as recent as possible.

A major problem with frequency counts is that they are based on the formal definition of words and they usually contain only single lexical item entries. For this reason, concepts that have a single meaning denoted by a group of words such as phrasal verbs, cannot be represented with this measure. Since the meaning that the words in an expression convey is unique and most likely distinct from the component words, it does not seem reasonable to form a combined, and a balanced frequency count to express the frequency of the expression.

2.6.6 Number of contexts

The number of possible contexts where the word can be met or used is important and is linked to the exposure to a word. If a word can be met in several contexts it is more likely that the learner has learnt the word. It is noteworthy that even though the learner may have met the word, he / she may not have learned all the senses connected to it.

When texts are compared to each other, they are shown to share some vocabulary while other words are specific to that text and that context. It would therefore be a reasonable assumption that exposure to a word met in several contexts would indicate that the learner has increased changes of meeting the word, even if that word may not be frequent.

2.6.7 Semantic grouping of English words

One possible classification of words is according to semantic fields. For instance, *table*, *spoon* and *kitchen* can be classified as *household* vocabulary. Such a categorisation is naturally subjective, and can be performed in several different ways by different people. If, e.g. a word like *lamp* is included in the household vocabulary, it can be argued that it could also occupy other categories, for example the category of *office* or *workplace*. One problem with such categorisation is to determine the level on which words can be grouped together, i.e. what are relevant and sufficiently broad categories to use in classifying the words. This problem becomes more prominent the more frequent the words in question are. Grammatical words like *there* and *he* can either occupy all the categories, or be listed separately in a general category. Words that have several different senses that are used in different contexts present a problem: which category should include these words. The problem is more difficult for open-ended test items, as in the multiple choice items a sense has been forced on the word.

Despite these problems, content categorisation could be useful. If language learners pursue language studies through communicative language teaching, it is likely that they study language primarily regarding those areas that they think they will find useful in the future. For this reason, their interest will guide the exposure to words. This exposure is of special interest for words that have a low frequency, and thus appear in fewer and more specific contexts. Categorisation is also important from the point of view of test bias. It may show if some areas are under- or over-represented in the test or if certain words can be associated more closely with different age groups, occupations, or either sex.

2.6.8 Basic and core words

Some words are more important for communication at different levels of proficiency. Even on lower levels of learning basic words are needed to carry out simple everyday needs and situations, such as grocery shopping or when asked for directions. However, even if it seems clear that a minimum vocabulary is important, it is problematic to define clearly what actual words should be included in that vocabulary.

As part of a research project to define minimum second language vocabulary (Basic English) Richards (1943) lists 850 words that were selected for beginning language learners because of their ease of learning. The selection of core words is based on a combinations of several types of information: frequency count of the words, importance for the learner etc. Carter and McCarthy (1995: 4) express criticism against this definition of basic vocabulary, because it is not clear how this vocabulary might be extended. Indeed, there seem to be no clear criteria for the selection of these words that would also be useful on higher levels of proficiency.

One possible detour to identifying core words is to use Carter's (1989: 33) tests of coreness. Properties of core words include syntactic substitution, antonymy, number of collocation relationships, extension, superordinateness, culture-free, use in summaries, association, neutrality of the field of discourse and tenor of discourse. Carter (1989: 44) states, however, that the research on coreness is still incomplete and advices to use caution, when deciding if a word is an actual core word.

2.6.9 Polysemy and homonymy

Polysemy refers to the different meanings or senses a word may have. Homonymy is used of a relationship where two words which have the same appearance have a different meaning. Vocabulary learners most likely become acquainted with polysemy and homonymy by looking up words in a dictionary, and discovering that a word may have several definitions pointing to different concepts.

It is possible that learners on the intermediate level focus learning efforts to word-sense relations only, but do not attempt to extend this knowledge far enough to other senses of polysemous words or homonyms. In this case the number of homonyms may be an insignificant factor in vocabulary learning, and the learner may be aware of the most commonest senses of words only. It is also possible that because of the ambiguous word-sense relationships, learners avoid using or learning certain words, because they are difficult to learn and a synonym can be used instead.

The multiplicity of meanings relates to coreness of words. Words which have several meanings are usually more frequently used, and this would suggest that language learners have better access to these words. However, there is not sufficient evidence either way indicating if these words are easy or difficult to learn.

2.6.10 Abstractness of the word

Schwanenflugel and Akin (1994: 251) define concrete words as those for which there are direct sensory referents and abstract words as those which do not. They also note that concreteness seems to represent a fundamental semantic distinction among words, in that concrete words seem to be more related to the image system, and abstract words to the verbal system of the dual-coding theory, i.e. that concrete words evoke images more easily than abstract words, but also that abstract words are more closely related to verbal expression rather than imageability.

The consensus among language teaching research is that words referring to concrete things are learned easily (see Carter 1989 and Nation 1990). They refer to things that can easily be illustrated by drawing, and it is thus easy to establish a one-to one relationship between a word and its meaning. Schwanenflugel and Akin (1994: 259) and Schwanenflugel and Stowe (1989: 122) found that concrete words require less processing time for elementary school children, but also for college students. They also noted that the processing of meaningful information was different between children and adults. Thus, the abstractness may be related more to the way children and adults understand the world around them. But it is also conceivable that adult learners on the first stages of learning find concrete words easier to learn. For example *pain* or *ache* are sufficiently important even on early stages of learning, and they are yet fairly abstract.

2.6.11 Imageability of the concept

Research has shown that imageability of the concept is an important factor affecting the ease of acquisition of the word. If a word rouses a mental image easily, it has a strong imageability. It is therefore easy to see that imageability is closely tied to abstractness of the word, as was noted above. However, the way mental images are roused is subjective. It is therefore difficult to perceive this quality as a very coherent structuring principle for vocabulary items. However, with reference to the above example it can also be said that *pain* for example may be relatively imageable concept, although it has an abstract meaning. Imageability has, however, been noted to relate to the difficulty of learning words (Ellis and Beaton (1993).

2.7 Measurement of vocabulary size

Measurement of vocabulary involves appreciation of the complexity related to the acquisition of vocabulary knowledge, and thus on what is to be measured. Spolsky (1989: 71) claims that proficiency measures are directed towards performance and use of language rather than competence of language. In criterion-referenced testing, the scores on a test are linked to ability descriptors or criteria which have a meaningful interpretation via sampled performances in simulated real-life situations (Bachman 1996: 61).

Measurement of vocabulary size is equally performance-based, but it is more closely related to competence. Bachman's model of language is a basis for several test batteries and it broadly divides language competence into organisational and pragmatical competence (Bachman 1990, 1996). Organisational competence is further divided into grammatical and textual competence, and vocabulary knowledge is part of the former. Vocabulary knowledge involves the knowledge of selecting and arranging appropriate words into sentences to express ideas. (Bachman 1990: 87). The model is not in disagreement with Nation's characterisations of word knowledge examined earlier in chapter 2.3. The vocabulary size can, in this model, be considered parallel to Nation's idea of translating ideas into actual words.

If this is the conceptual place of vocabulary size in a language system, what is then the appropriate instrument with which the it can be measured? Meara (1989: 66) says it is not clear which method is most useful for vocabulary size measurement. For this reason several methods exist, and it might seem reasonable to say that one task type does not necessarily tell very much about the control or magnitude of the size of vocabulary. But if several task types are used, problems have to be faced concerning the interaction of these task types in producing an estimate of proficiency. However, before examining that interaction, it is important to understand how individual task types function within a test battery.

Measurement is based on the assumption that the measure used should focus on one single trait or skill (unidimensionality) and the responses to items should also be independent of each other (local independence). Several formats have been developed; some of them are better suited for the measurement of vocabulary size than others. Read (1993: 356) describes three continua on which task types can be mapped. The first is verifiable responses vs. self-report. It seems evident that for certification purposes verifiable response tasks would be preferable. The second continuum is testing in context vs. in isolation. Use of context has been proposed in order to enhance the authenticity of the tasks. The use of context, however, increases the examinees' processing time of the task, and Goulden et al. (1990) see this as uneconomical for testing, because a relatively small number of items can be processed in comparison to context-free items. It is also difficult to estimate what amount of context is sufficient for preserving the communicative authenticity of a task. Use of context may also affect the unidimensionality assumption of tasks in that authentic tasks usually require a combination of skills to be used for a successful response to be produced. Testing breadth or depth is Read's third continuum. Tasks may attempt to find out how many words a person knows, or how well he knows them. The fact that a learner has a large vocabulary does not necessarily mean that he can understand the words when met in context, or that he is able to use them appropriately in real life situations.

2.7.1 Cloze and C-tests

Cloze and c-tests are tasks in which the examinee is presented with a text from which a number of words are missing. The words can be deleted by a random selection method (e.g. taking every *n*th word out of the text), or by selecting specific words for deletion. In receptive skills testing, the examinee is to supply the missing words by selecting from a set of alternatives.

Complete the gaps in the text below.

The matter of predictive validity is something the examining boards really ought not to _____ concerned with. Examinations index achievement over a specified period of ____ and, as such, are essentially backward looking.

FIGURE 2.1. An example of a cloze task.

Chapelle studied the use of c-tests in L2 vocabulary research. She (1990: 177) found that the tests should be used carefully, because e.g. random selection of missing words may turn some items into grammar items. The results also cannot necessarily be attributed to vocabulary size alone. This would seem to indicate that c-tests are better suited for the assessment of the examinees' ability to structure words into coherent texts.

2.7.2 Yes-no task

The yes/no task is composed completely of target language words. Usually a small proportion of imaginary words are added. These, however, should look like the target language words. The examinee marks for each word whether he thinks it is a real word in the target language. The advantages of this test types include ease of item construction, inclusion of a large number of items, ease for low ability examinees to answer and a potentially high correlation with reading ability (Anderson and Freebody 1983, Meara and Buxton 1987).

Shillaw investigated the properties of the yes/no task. The task was used in different compositions and settings, and the results show that the reliability (KR-20) did not exceed .848 in any of the testing conditions. The best reliability was achieved when no imaginary words were used. The test was also clearly unidimensional, because the first factor in factor analysis was markedly higher when compared to the second factor. (Shillaw 1996:4). Shillaw also noted that

In the following you will see a list of words. Mark 'yes', if you know this word is a real word in English. Mark 'no', if you think it isn't.

	yes	no
rail	□	□
snat	□	□

Figure 2.2. Two examples of the yes/no task items for a real and an imaginary word.

when non-words were used, these led to misfit of the non-word items. For misfitting real words, guessing and poor discrimination are suggested as a reason. (Shillaw 1996: 5). Shillaw, however, reports only moderate correlations (below 0.5 sig. .000) between the yes/no task and a proficiency test administered to the subjects.

The task type, however, has problems with the scoring procedure. When imaginary words are used, it is possible that the examinee may in some cases receive a negative score on the test, and the administration becomes more problematic in large scale testing. Shillaw (1996: 8) recommends using Rasch-based scaling as a method for solving the scoring problem, which also enables the comparison of test scores to other proficiency criteria. Inventing imaginary words is not always easy, as it is possible to come up with a word that is used in a dialect or vernacular of the language. One way to deal with this is to rule out the imaginary words completely, which does not seem to weaken the test (Shillaw 1996: 7), but makes it less useful for high-stakes testing.

2.7.3 Nation's Levels test

Nation proposes a task type in which the examinee is presented with items that consist of six words and three clues that correspond to those words, e.g. their definitions. The examinee then has to find a match between the word and the definition. This task type is easy to mark, has low rate of guessing correctly, economical (large numbers of words can be tested), and allows learners to use different kinds of knowledge of the words that they have (Nation 1990: 261).

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning.

1. business
2. clock ______ part of a house
3. horse ______ animal with four legs
4. pencil ______ something used for writing
5. shoe
6. wall

(Source: Nation 1990:264)

FIGURE 2.3 An example of the Nation's Levels test item.

A task described above is actually composed of three items, for the examinee has to have all three matches correct in order to score full marks on the task. In testing this would correspond to three items. In measurement terms this task would be referred to as item bundles (Rosenbaum 1988: 349). This means that all the words in one task have to be assessed as a group of three items, and the information about any one of the items is lost. This means that the individual words should be as similar from the measurement purposes as possible, in order to get accurate and useful information about the size of the examinees' vocabulary. Furthermore, it is difficult to construct items for this task type and to make useful combinations of words for the items.

2.7.4 Multiple choice tasks

Multiple choice tasks can be realised in several formats. The tasks are based on the general format with a stem and distractors. Stem refers to the stimulus or problem given for the examinee, and distractors are the alternatives from which the learner is to pick out the one he thinks is the response to the problem presented in the stem. The tasks usually differ in the kinds of problems presented for the learner, or in the number and kinds of responses given to the examinee.

Choose the word that you think is the best translation for the English word.

sailor a) sairaala b) purjehdus c) merimies d) kuorma

FIGURE 2.4. An example of a multiple choice item.

The stem can be presented in several ways. The stem may contain a picture, text or spoken material to present the problem to the examinee. Until recent development with computer based testing, visual and aural materials have been more difficult and expensive to prepare, which means that they have not been favoured in test construction. Textual media have been preferred in vocabulary size studies also because they are easy and fast to administer, and thus a better range of the vocabularies can be covered. Text-based stems can present the problem in different amount of contextual information. A single target word is to be preferred (see figure 2.4), because in this way the knowledge of a word is restricted to one word only, but the assessment is context-free and does not require the examinee to understand the word in a context in which the examinee might not be able to distinguish the correct response. This also simulates a situation that happens in reading, when the readers attention is focused upon a word for some reason, for example when he is unfamiliar with it. It also helps to avoid the problems related to content sampling. The word and the meaning become independent of any specific context area, but the fact that the item focuses the examinees attention on one single meaning-form relationship is also the problem with which a reader is faced with in reading.

The number and quality of alternatives is crucial for a successful task construction. Too few alternatives increase the examinee's chances of guessing the correct alternative, while too many of them may make the task complicated.

The chances for guessing the correct response for a two-distractor item is 50%, and drops down to 25% for four-distractor items. Four distractors have been used in the Finnish National Certificate examinations (hence FNC) tasks to keep the tasks manageable and relatively fast to complete by the examinees. Pressley and Ghatala (1988: 462) investigated the use of multiple choice tasks for reading comprehension in connection to other task types using confidence ratings of the examinees to estimate how sure they were about the answers that they gave. A noticeable result was that some learners tended to be more overconfident about incorrect responses to multiple choice items as being correct than on other task types. They concluded that this overconfidence may be an indication that the examinees do not process the items adequately once the illusion is there.

From the linguistic point of view the multiple choice task type has been criticised in that it does not adequately test the competence of a language learner (for summary see e.g. Wood 1991: 33). The fact that the examinee has to choose from a set of fixed alternatives is probably the most notable drawback of multiple choice tasks. It can be argued that a low number of alternatives do not necessarily correspond to those that the learner would have in mind when faced with the word in real life.

For certification purposes the multiple choice task seems best, because it represents a task type with verifiable responses, is relatively fast and economical to administer, and seems to focus on the receptive knowledge of meaning. The measurement quality is also more direct, in that inferences between a word and knowledge of its meaning can be more clearly and objectively demonstrated, than in the other task types.

2.8 Aspects of multiple choice items

The difficulty of ~~an~~ the multiple choice item can, in this case, be seen as the function of all the components of that item (starting from the instructions) and their interaction. There are several steps involved in the construction of a multiple choice test. These include specifying what is to be tested and how instruments are constructed.

2.8.1 Specifications

In order to solve some of the problems of the multiple choice items, task construction should follow agreed upon principles or specifications. Specifications should give a detailed account of what is to be tested and how testing should be conducted. It is, however, difficult to identify features that would enable efficient task construction prior to the administration of the items. For example, the specifications for vocabulary testing in the National Certificate Examinations focus on the vocabulary size estimation by describing the skills assessed (i.e. vocabulary size in this case), a selection / sampling procedure of words, description of task types with examples and the scoring of items (Yleiset Kielitutkinnot 1995: 13).

The basic task type used in the assessment of knowledge of word meanings is the multiple choice (Yleiset Kielitutkinnot 1995: 13). A bilingual task is used on both the basic and intermediate levels. The test on the intermediate level should focus on a vocabulary of 5,000 most frequent words. Even though the item type is described on the surface, a detailed specification of item construction has not been specified. An important part in constructing multiple choice items is to pay sufficient amount of attention to the relationship between the correct answer and the distractors. Characterisations of what are perceived to be good items are usually general such as those in the U.S. Department of Health, Education and Welfare (1974: 8) study of vocabulary size tests.

> 1. The distractors [are] less difficult than the stem word and at the same or slightly lower difficulty level than the correct response.
> 2. The distractors [are] in parallel form to the stem word, the correct response, and each other in regard to tense and part of speech.
> 3. Spelling and sound similarities [are] avoided between the stem word and the distractors except where necessitated because of sound or spelling similarities between the stem word and the correct response.
> 4. Distractors [are] chosen to assure that they had no relationship to any of the definitions of the stem word.
> 5. Effort [is] made to keep repetition of distractors (and correct response) to a minimum throughout the test.

Also Herman (1988: 359) lists similar properties of good items. He lists 17 general features that items should have emphasizing rational approaches to be used in what to include in the item and how to best distribute them into the stem and distractors.

2.8.2 Vocabulary sampling

Sampling is not related to multiple choice alone, but is a step in the process of constructing any of the above methods. The basic idea is to estimate vocabulary size with reference to a representative sample of words. Sampling involves defining a sampling frame which lists the domain of words included into the definition of word. In practical terms, the sampling frame is usually a dictionary or a word list. After a suitable source material in chosen, the usual process is to estimate the number of words in a dictionary, obtain a random sample of words following the definition of what words are to be tested. Tasks are prepared for the examinee, and, after administration, the learners vocabulary size is calculated from the proportion of words he knows of the sampled words by relating it to the total number of words in the sampling frame. At least two crucial steps are involved in this process: specifying the criteria for selecting words into the sample, and establishing a reliable testing method to assess the knowledge of these words. (Nation 1993: 27). Problems in the adequacy of sampling, thus, may lead to test results from which it is not useful to estimate the size of the learner's vocabulary.

The Finnish National Examinations states that the number of tasks should be kept to about 40 in order to keep the test manageable. The target words for these items are preferably sampled from a dictionary, because vocabulary lists based on frequency counts do not usually represent the spoken language in an adequate manner, and they are not based on language necessary for language learners. Even though a frequency list should not be used directly, it is advisable to estimate the level of the words if reference to frequency information is available. A useful source has therefore been deemed a more limited dictionary of 10,000 words, from which the target words are sampled. The test constructors are then asked to remove too easy or difficult words from this sample. (Yleiset Kielitutkinnot 1995: 13).

2.8.3 Error categorisations and the multiple choice items

Distractors are the incorrect options of the multiple choice item. In order for them to be plausible so that the examinee might be tempted to select one of them when he / she does not know the correct answer, it is usual to design them according to errors that examinees make. These can be collected, for example, with the help of open-ended items or from compositions.

There are different kinds of divisions of learner errors. A classification can be drawn based on the research on the structure of the learner lexicon. Earlier studies distinguish between learner errors based on the arrangement of the lexicon (Henning, 1973, Nation 1990: 35). Above it was noted that the learners tend to organise lexicon according to formal / aural principles in the early stages of learning, and according to semantic principles later on. A broad categorisation would probably categorise errors into those based on sound and form, and those based on semantic aspects.

The items themselves have to be quantified in a sensible way to see to what extent the facility of the item itself affects knowledge of the word. A categorisation based on the structure of learner lexicon (hence classification 1) observed in the present items is shown in table 2.3. The basic division in the first classification is between semantic features, appearance and sound, or general plausibility. In the first category all distractors included referred to semantic errors. For instance if the distractor referred to a concept from the same semantic group, or a word that was a hyponym of the correct word. Also included were words in which part of the word referred to a different concept. In the second category all syntactic or phonological similarities were included. All words with no conceivable syntactic or semantic relationship, except for plausibility, as a distractor were included.

A classification of psycholinguistic sources of errors shown in Ellis (1994: 58) is presented in figure 2.2. Error sources can be divided into two major categories, competence and performance error. Competence errors result from lack of knowledge in the target language, while performance errors are involved in problems related to communication situations where situational factors may overload the language learner's capacity to perform in a situation and hence cause erroneous language use. For this reason Ellis refers to performance errors as mistakes rather than actual errors.

This classification was applied to the items in the present study. A classification of performance errors was not included, as this would require additional information from the test subjects, which was unavailable for this study. Competence errors were divided into the three categories shown in figure 2.2, and the distractors were labelled accordingly.

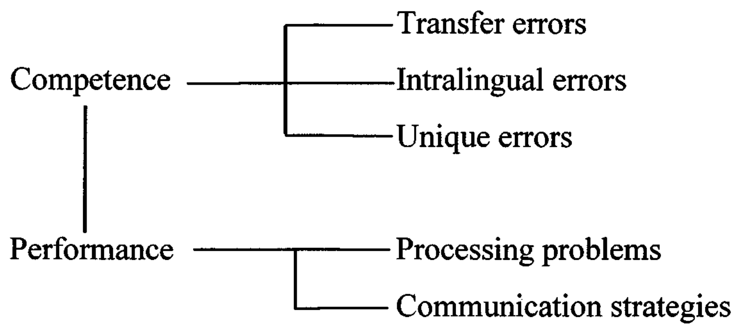


FIGURE 2.2. Psycholinguistic sources of errors (from Ellis).

The second classification had a slightly different orientation to the item structure. Under the first category, all transfer errors can be included. This includes errors arising from similarities between, for example, Finnish and English. Also, similarities resulting from other languages, such as Swedish can be included into this category, because Finland has two official languages, and the learning of Swedish is mandatory. The second category includes all errors that result from inaccurate use of the English language system. Unique errors are errors that do not have a linguistic explanation. Distractor words that refer to a meaning relationship that is not clear can be included in this category.

A. Classification based on the structure of learner lexicon

1. Semantic features

- Hyponomy (desk: pöytä - *laatikko | drawer)
- Same semantic field (riddle: arvoitus - *arvaus | a guess)
- Concept overlap with part of the word (foam: vaahto - *muovi | plastic)
- Deceptive transparency (income: tulot - *aula | lobby)

2. Appearance or sound

- Similarity of sound between target word and Finnish alternative (immoderate: kohtuuton - *modeemiton | one without a modem)
- Similarity of appearance of target word and a translation of a distractor (association: yhdistys - *syytös | accusation)
- Similarity of appearance or sound between a target word and a word from a third language (grease: rasva - *possu | piggy / SWE gris)
- Morphology: word part suggests meaning (e.g. -less) (gutless: pelkuri - *avuton | helpless)

3. Plausibility

(molar: poskihammas - *valas | whale)

TABLE 2.1a. Distractor categorisations based on the hypothesised structure of learner lexicon.

B. Classification based on psycholinguistic sources of error

1. Transfer

- Distractors based on confusion arising from a similarity between in the English word and a Finnish / Swedish word
(grease: rasva - *possu | piggy / SWE gris)

2. Intralingual

- Distractors based on incorrect use of the foreign language system
(pond : lampi - *osake | bond)

3. Unique

- Distractors based on other error types
(dandruff: hilse - *keimailija | flirt)

TABLE 2.1b. Distractor categorisations based on psycholinguistic sources of error.

2.8.4 Other item features

Another possible way to measure item performance is to see how many unattractive distractors were included. A commonly used criterion for a useful distractor is that over 5% of the examinees should choose it instead of the correct alternative, and this criterion was used also in this study.

It is also possible that ~~the~~ distractors do not work well for the different levels of proficiency. It seems reasonable to assume that learners on lower levels of proficiency are tempted by distractors more than higher level learners. There can be several reasons why this happens, one of them can be the fact that one of the distractors is too close to the correct choice, and could also be regarded as a correct response. One way of assessing proper distractor functioning, is to divide the sample of examinees into subgroups according to their ability (measured e.g. with their total score), and to plot a curve of the proportion of examinees selecting that distractor in the different ability levels. When items work as they should, the line for distractors starts from a high value indicating ~~a~~ the number of persons selecting the wrong answer, and it moves progressively to lower values as the proficiency increases and the examinees select the correct response. The opposite, of course, applies for the correct response. When the line turns out to be different, there may be a reason to suspect that the item encourages examinees to good guessing.

2.9 Item parameters

There are two parameters that are important for describing the usefulness of items: item difficulty and discriminability (Konttinen 1980: 62 and Wood 1998: 377). Item difficulty is an important characteristic used in the scaling of test items. In order to economically design tests, item difficulty should be kept close to the level of the examinee. Too easy or difficult items will not give him the changes to show the ability in question. The FNC examinations are administered for three ability levels, from which the examinee can choose the one most likely matching his level. It is therefore useful to know which items to administer for these levels prior to assessment.

Item difficulty can be defined prior to the administration of the items only through item characteristics such as the word features described above. This, however, cannot predict definitely how the items will work in the test. The most efficient way to estimate item difficulty is to define it in relation to the responses produced by examinees. For this reason, items are usually pretested on a representative group of examinees in order to produce estimates of item difficulty. These estimates can be used in the construction of the actual test, when the difficulty estimates are relatively stable from one administration of the test to another. Currently there are at least two widely known methods for estimating item difficulty: classical test theory and item response theory.

In classical test theory item difficulty is defined as the propotion of all correct responses to an item (see e.g. Konttinen 1980: 62). It is called the p-value. This index actually provides the facility of items, i.e. how easy the items were, because a high value of the index means that the item is easy. For this reason, the inverse of p-value is used to tell how many people got the item wrong in the test. The classical difficulty index is, however, sample-dependent and can be used as a rough guide for generalisations of the difficulty of items to further populations of examinees.

The problem of sample-dependency has been overcome in item response theory, where an item's difficulty is defined as the log-odds of the examinee to succeed in responding to it. The advantage of the logarithmic scale is that it provides a yardstick stretching to infinity at both ends, but which shows the relative distances between items in terms of their difficulty. On this scale the difficulty of an item is expressed as a meaningful distance to other items. The scale also provides information of the person taking the test, the person's ability is expressed in a similar way. Also, the scale does not have an absolute zero-point, but it has been customary the mid-point of the scale into zero. In this point both the ability and difficulty of items is .5, which means that a candidate with the ability estimate has a 50% probability for producing a correct answer.

The difficulty estimate alone, however, does not tell enough about the usefulness of the items, as it is possible that the item does not validly measure the concept: it is possible that examinees who have low ability get the item correct while the examinees, who have high ability get the item wrong. This is possible for example in cases where two of the distractors in a multiple choice item could be considered the correct alternative or if low-ability examinees can easily guess the correct response. In this case the item does not discriminate well between low and high ability. Usually, a biserial or point-biserial correlation is used to represent this index. Both of them calculate the correlation between an item and the total score of the test.

Estimation methods have been extended to two and three parameter models that take into account also the discriminability and guessing involved in the responses. The advantage of one-parameter model is that it does not requires complex computation of parameters or large samples to be obtained. However, it is useful to have at least 200-300 candidates to calculate sufficiently stable estimates, but also an excessive number of examinees may result in poor estimates (Hambleton 1993:172).

An additional advantage of the item response theory based estimation of item difficulty is the way items get a measure of error associated with each of them. In other words, it is possible say how well the items fit on the model that expected the vocabulary size items to measure vocabulary size, and thus have a reliability measure related to each item, but also for each test taker (McNamara 1996: 132). Classical test theory does not provide an estimate of error related to item difficulty, instead a reliability index for the whole test is calculated. One way to estimate the error related to items in this framework is to see how test reliability changes, if the item was not in the test.

There are several ways to compute the item parameters for test data. The American measurement tradition has developed software for this purpose (FACETS; Wright, Linacre and Schulz 1990 and BIGSCALE; Linacre 1994), which is based on the strong assumption of IRT. It is assumed that the items used in a test are equally powerful in distinguishing ~~the~~ low proficiency examinees from those with higher proficiency, and there is evidence that this may not be the case with that many tests (Hambleton 1985: 46). The estimation process should therefore take the discrimination into account. For this reason, another computer application was used to compute the difficulty parameters for this study, namely the OPLM, or one-parameter logistic model (Verhelst, Glas and Verstralen, 1995). This program is based on the model described above, but it treats also discrimination as something that has an effect on the parameters by using conditional maximum-likelihood estimation method (Verhelst, Glas and Verstrahlen, 1995: 2 and Hambleton 1985: 81), and thus the model can accommodate also items that do not discriminate equally well.

OPLM also provides an index of the global fit of the items to the unidimensional model. The items that have a poor fit, i.e. measure too much error or other abilities, will decrease the global fit, and therefore these are usually taken out of the subsequent administrations to achieve better fit and thus better measurement quality.

2.10 Scaling

The object of scaling is to define an instrument with which it is possible to measure a concept in different points of time. Defays (1988: 316) defines scaling as ”establishing a correspondence between a set of data, with observed relations, and a set of numbers.” In other words, scaling means measuring attributes of a group of cases and stating what the relationship between the cases is according to the attributes. The relationship can, for example, be height, in which case we can state the relationship of two persons, but it can also be a psychological characteristic such as language ability. These relationships are usually expressed in distances between objects (Dunn-Rankin 1988: 307).

The data in Defay's definition can come from a variety of sources and thus it can be measured on different types of scales according to the information that the scales include. The Most common scales with the educational research are the nominal, ordinal and interval scales. Nominal scales include information about the class of a case, e.g. the mother tongue of an examinee. An Ordinal scale expresses the rank of cases, as for example the total score on a test orders examinees according to their total score on the test. An Interval scale expresses information about the relative distances between cases. For example temperature is measured on an interval scale. All these scales have the properties of the preceding scales, but the opposite does not apply.

An important property of a scale is that it can be applied from one situation and circumstances to another. Scaling of human activity is more problematic than for example in the natural sciences. Rasch (1980: 10-11) compares models of classical physics and psychology to each other and notes that the main difference is in the fact that classical physics can formulate laws which can be used to predict the movement of objects, whereas human behaviour has to take into account the infinite variability of human action. It is virtually impossible to predict how a person will succeed with any given task in advance. Thus, Rasch (1980: 11) recommends the use probabilities instead.

In educational measurement unidimensional scales are preferred. This is because of the simplicity inherent in the interpretation of the results (McIver and Carmines 1981: 15). As already noted, a unidimensional scale measures a single property. It is, however, not always that easy to distinguish this single skill in measures of human behaviour. Dunn-Rankin (1988: 309) mentions four methods of unidimensional scaling, of which Guttman scaling is the most useful for the measurement of ability. This scaling method arranges test items into an order of difficulty, based on response patterns. The scaling method has problems with the selection of measurement instruments or test items, because for the scale to be useful a careful selection of useful items is required. (Dunn-Rankin 1988: 312). These limitations arise mostly from the sample-dependency of measures used in Guttman scaling. It is not useful to construct tests based on such measures, because subsequent administrations may refute the scale. Rasch's alternative was to use probabilities as measures, which enables measures to be more precise between administrations of a test.

With regards to IRT the assumption of unidimensionality has brought more controversy in using statistical models with language tests (see e.g. Buck 1994 and Henning 1988). For the assessment of communicative language skills, for example listening, Buck (1994) considers it difficult to assume that a unidimensional model could be accommodated for performance data on listening test items. McNamara (1996: 296) notes that the skill in itself does not necessarily have to be such that people linearly acquire more and more difficulty aspects (i.e. the skill is not unidimensional), but the measurement instrument should make up test of items that can be said to measure ~~a~~ the concept in a way that would provide a progression from easy to difficult items in terms of one ability.

Multidimensional scaling methods are more useful scaling methods, because they can be utilised with human measurement with more precision. For example, in the measurement of language ability there have been problems in specifying what the focus of measurement should be. Merril and Swain (1980) provided the first operationalised model of communicative language ability which has been added to subsequently by Bachman (1990) and others, but it is still difficult to say with absolute certainty for instance, which aspects truly belong to grammar and which to vocabulary. Therefore, in a testing situation, it is not always clear how to make a distinction between the competencies that are tested. This is evident in languages (for example Finnish) in which it is not possible to assess vocabulary without testing also grammar at the same time. In these situations multidimensional scaling methods would seem to be justified.

McIver and Carmines (1981: 13) point out that although multidimensional scales are likely to be more precise, they are more complex to interpret. They would therefore prefer the use of unidimensional scales in the measurement of psychological abilities. They also point out that multidimensional scaling methods may be used in the construction of a unidimensional scale, but not vice versa. Dunn-Rankin (1988: 313) mentions two useful multidimensional scaling methods: factor analysis and multi-dimensional scaling. These can be applied to unidimensional scaling, by focusing the analysis only on a single dimension. The idea is to produce a scale that describes item performance on several dimensions at the same time. It would therefore seem to be important from the point of view of test validity to examine a data set by using different kinds of scaling methods.

The importance of scaling is that it enables valid test administration, and more importantly valid inferences to be made based on the test. If a test is to be considered unidimensional, it is necessary to make sure that, at least on the whole, the test is mainly measuring a single skill. If a test that is thought to be unidimensional seems to measure several skills simultaneously, it is necessary to investigate what aspects of the items cause this.

3. METHOD

3.1. Research design

The main purpose of this study was to investigate the principles that can be used in a scale of a vocabulary size measure from a multiple choice task. The specific research questions have been shown in table 3.1. On a general level the goals can be summarised under three main areas. Firstly, test quality is assessed in order to estimate the adequacy of the test as a whole for scale construction purposes. Attention is here paid on the test reliability, fairness and validity, item parameters, and model-data fit. Secondly, test structure and dimensionality of the test was investigated in order to establish what exactly was being measured. Thirdly, and most importantly, the relationship between content-based item descriptors and empirically-based item characteristics were compared. Adequate item performance necessitates that items cover the subject area adequately, and do not cluster around one or a few attributes. More specific research questions, the methods used and the corresponding results sections are listed in table 3.1.

3.2 Two data sets used in the study

In order to address the research questions it was necessary to obtain two sets of data. Both of these are based on the test instrument and they can be divided into two groups: a priori features of the target words, and items and a posteriori item characteristics. The first group of features were assigned to items based on features of the items by reference to such features as word class etc, and this constituted the first data set. Some of the measures in the first data set represented subjective aspects of the words, and for this reason raters were used to assess the features. Six raters took part in the rating process and the median of the ratings

Main area	Research question	Method	Results
1 what?	1. a. How are semantic fields represented in this study? 1. b. How are word classes represented in this study?	- Pearson correlation	4.1.1
1	2. Is the test reliable?	- Cronbach's alpha - p-value in groups	4.1.2
1	3. Is the test biased towards either sex or age groups?	- p-value in groups	4.1.3
1	4. What does the item difficulty scale look like?	- p-value - b-parameter	4.1.4
1	5. Does the data fit the Rasch model?	- R1c test	4.1.5
1	6. Do the items discriminate between groups of low and high ability examinees?	- point-biserial correlation - biserial correlation	4.1.6
2 what?	7. a. What is the internal structure of item variables? 7. b. What is the relationship of the frequency counts?	- Pearson correlation	4.2.1 4.2.2
2	6. How many dimensions are identified?	- inter-item correlations - scree-plot of eigenvalues - stress-test	4.2.4 4.2.5 4.2.6
2	7. Do the items form clusters?	- multidimensional scaling - hierarchial clustering	4.2.7
3	How are the content-based characteristics connected with item difficulty and item discrimination?	- Pearson correlation	4.3.1
3	8. Can the dimension(s) be linked to word and item features?	- Pearson correlation of principal components and item variables - Pearson correlation of dimensions in multidimensional scaling and item variables	4.3.2 4.3.3
3	9. Can the word and item features be used to explain the clusters formed from the performance data?	Discriminant analysis	4.3.4

TABLE 3.1. Research questions addressed in this study.

was used as the measure of the rated features. The second group of features were assigned to items after administration of the set of multiple choice items, when it was possible to compute such indexes as empirical item difficulty and discriminability etc. These indexes made up the second data set.

3.3 Instrument

A test of vocabulary size used in FNC English examinations on the intermediate level, spring 1996 was used as the instrument in this study. This was the third year of administration of the examinations, and the items were constructed with minimal specifications. The main requirements have been examined in 2.11.1, when the specifications were discussed. The test, thus, consisted of 40 multiple choice items

constructed for randomly selected target words. An example of an item can be seen in figure 3.2, and the whole test in appendix 1.

M28. Riddle a. ristisana b. arvoitus c. palapeli d. arvaus

FIGURE 3.2. An Example the multiple choice tasks used in the study.

Finland has two official languages, and consequently the assessment system must, by law, administer a Swedish version of the test for those examinees who wish to take the test in Swedish. For this reason the items were translated into Swedish. Usually the number of Swedish speaking test takers is rather low, and on this session eight candidates took the tasks in Swedish. Performance data was coded according to the conventions for multiple choice tasks: correct answers with 1 and incorrect answers with 0. Missing answers were coded as blank. For the purposes of this study, these items are coded with zero.

The definition of word was based on the communicative competence model of language, and hence the emphasis was on assessing the number of word-form - meaning relationships the examinee knows. The assessment procedure follows that described by Nation (1993), and the sampling of words was performed by randomly selecting 40 words from a translation dictionary of 10,000 words. The only limit in accepting words into the sample was that it was not a proper name or potentially offensive for the examinees. Items were constructed as four-alternative item multiple choice items, and in constructing the items attention was paid to the general rules described for multiple choice item construction and keeping the distractors as plausible as possible. The test is shown in appendix 8.

3.4 Subjects

The examinees took the test as part of a test battery designed to assess their language proficiency in five sections: reading, writing, listening, speaking and vocabulary / structures. A total of 475 examinees with different kinds of backgrounds took this task. The sample represents a typical sample of candidates taking part in the FNC examinations in that the representation of people from different walks of life, age groups and sexes were present. However, this administration session contained an unusually large number of examinees.

The subjects' age ranged between 18 and 64 years with 293 women and 182 men. In order to assess the test fairness, the subjects were divided into groups according to these attributes, as shown in table 3.3. The three age groups were combined from an original number of six groups because the representation of the youngest and oldest persons was not as high.

	Women	**Men**	**Total**
Age 18-28	91	60	151
Age 29-41	134	88	222
Age 42-64	68	34	102
Total	293	182	475

TABLE 3.3. Group sizes for age and sex.

3.5 Variables used in this study

The features included in this study are shown in appendix 3. Figures in brackets indicate the alpha reliability of the ratings. Some of the item features were test statistics described above. One of the main reasons for selecting these particular features was the relatively easy availability or computability of variables. Such variables are helpful in the specification and construction of items, but research also indicates that they have an important relationship to the size of vocabulary. The variables have been ordered into two sets. The first set contains content-based variables of the items. The second set contains empirically-based variables.

The first content-based variable, abstractness, was entered as a dichotomous variable. If the target word in the item was abstract it was coded as one, but if it was concrete it was coded as zero. The differences in age groups were computed using the p-value. The p-values for items in the three age groups were calculated, and they were subtracted from the groups of the previous groups. Thus, negative values indicate that young test subjects did worse in the item and better, if the difference was positive.

The impact of affixes on knowledge of words was used as one index. The affixes listed in Quirk and Greenbaum (1973: 430) were used as a basis for identifying the affixes. A word was classified as a base word, when all these affixes were removed. Thus, the distance to the base word was computed as the number of affixes removed to the base word.

Basic words were computed as a binary variable. For this classification the list in Marzano and Marzano (1988) was used. If the target word appeared in their list of target words, the word was labelled as a basic word. Otherwise it was considered to be a non-basic word.

The number of contexts in which the word usually appears in texts was computed as one variable. The number of contexts was available from the LOB corpus and the numbers from that source were used. The LOB corpus contains 15 general context areas in which the word may appear (Johansson and Hofland

1989: 2).

Similarity between word forms was scaled on a three-step scale. The classification was done by the author. The written forms of the words were compared because of the written medium used in the test. It seemed also difficult to use the number of matching letters or syllables as a measure, so instead the appearance of similarity was used in the rating.

Three frequency counts were used. This was done in order to assess the usefulness of word counts based on sources from different dates and different sized corpuses. The first source used for this study is the British National Corpus lemmatised word list (Collins Cobuild 1996). It contains 49,999 uninflected words taken from a sample of a corpus of over 100 million words. It is based on both written and spoken source material with 7 times as much written material as spoken. (Kilgarriff 1996: 1). The second source, Frequency analysis of English vocabulary and grammar, is based on the Lancaster-Oslo / Bergen corpus (Johansson and Hofland 1989). It is based only on written material collected from British books, newspapers, periodicals and government documents printed between 1960-4. For the BNC a second index was computed indicating the rank in the frequency count. Also, a Finnish frequency count was used to establish the frequency of the Finnish words begin tested to see if the frequency of the concept in mother tongue had a significant effect.

Imageability of a word was rated on a three-step scale. Six raters were used, because of the small sample. The median of the responses was used as a variable. A similar rating process was arranged for the assessment of the importance of the words for learner. A three step rating scale was used, and the same raters were instructed to assess, how important they thought it would be for a learner on the intermediate level to know the word in question.

Numbers of distractors classified in the two classifications discussed above were calculated. Each class was entered as a single variable and the number of corresponding distractors in each item were counted totalling six variables. These classifications were thought to be interesting firstly to find out, if a certain type of error-based distractors consistently predict success in the items. Secondly, it was considered important to see if it makes any difference to use only one type of distractors in an item, or if a mixture of distractor types are more useful. It was possible to use these classifications for the present items, but it is also important to note that the actual items were not actually constructed with this classification in mind.

The number of homonyms was counted and the count was used as a variable, and the number of letters was counted in the same way to estimate word length. The use of different types of distractors was computed using both distractor error classifications. The difficulty index for item response theory-based b-parameter was computed with OPLM. The related error in the estimation of item difficulty was also entered as one variable.

Word class was entered as a group to which the word belongs, and, lastly, semantic categorisation of words into semantic fields was attempted. This classification is subjective, and thus it was decided to compare the existing categorisation to see how uniform the classification is. Three books were selected and compared, each listing central vocabulary of English for Finnish learners: Words Onwards (Avokari et al., 1980), Nyky-Englannin keskeinen sanasto (Miettinen and Uotinen, 1983) and Word Files (Avokari et al., 1997). A slightly different categorisation system has been used in all three books in terms of depth of classification: the second book clusters words into broad categories, while the third book employs the most fine-graded system. Two of the books are based on frequency, and the words have been divided into frequency groups to help in teaching. WOW contains 7,000 most frequent words, and EKS contains over 4,000 most frequent words. For WF this information was not available. WF contain the most up-to-date content areas of the three, and has the most elaborate system for classifying words. The content categories are described in appendix 1. The categories for each book are marked with number coding, but they are not exchangeable between different books, because of differences of categorisation systems used.

Each of these books has been directed at language learners wishing to improve their English vocabulary, and therefore the word lists are topic-driven in the sense that they mostly contain nouns, verbs or adjectives pertaining to each particular category. For this reason, it was not possible to find exact matches for phrases, and some words were missing from the categorisations altogether. Also, some words occupied more categories than one. However, these subject lists are likely to contain words that are a part of several courses and therefore are likely to represent words that learners also find useful and central for effective communication.

It seems that it is not possible to rely on any one source of categorisation for words, and for this reason, a fourth category was added summing up the available information from all different sources. Appendix 6 shows the diagrammatic categorisation of target words. Categorisation was also amended where deemed necessary. In these categories all word attributes were considered in labelling a category / categories for a word, but the meaning of the word was the main classification criterion. Also, where possible, the most frequent meaning of the word was checked in Collins Cobuild dictionary of English, and used in determining the category of a word. This attempt was made to establish a primary meaning and thus a single primary category for each word. It is also important to note that while these categories might be useful for this study, the classification is very subjective and context-bound, and most likely will have to be done again for other uses.

This classification was not clear for frequent or general words. A word can be placed in more than one group by semantic division, i.e. 'foam' can be in the semantic group of foam that comes from the mouth of a dog, but also in the group of plastic or rubber products (i.e. should only one semantic group be chosen for making the distractors or not?).

The second set of variables contain the empirically-based item chracteristics. The reliability related to each of the items was estimated through alpha reliability. It is possible to compute alpha, if that item was removed from the test, and thus show the effect on reliability of the test, if that task was not part of the test. As a measure of this change, an index was calculated, in which the changed reliability figure was subtracted from the observed alpha for the whole test. The resulting measure indicates how useful it would have been for reliability to either include of exclude the item from the test. Negative values indicate improved reliability and positive values decreased reliability.

Inverse p-value was used as an index of classical item difficulty. First the proportion of correct for each item was calculated, and then this figure was subtracted from one to get the value used in the variable.

Even though it was possible to apply the error classification for the present distractors, it is possible to take into consideration the attractiveness of the classification in order to assess the usefulness of different types of distractors. For example, the distractors constructed may be too easy for intermediate level learners, and thus the problem does not lie in the classification, but in the quality of the distractors. For this reason, it does not seem justified to calculate only the number of times each category was used, but it seems better to include a weight for each distractor strategy used. This was operationalised by calculating the proportion of responses for each distractor. The figures were then summed for each type used in the item, for example, if two distractors in an item were based on form, then the corresponding proportions were summed. Since the weight can be calculated only after the test has been administered, and it may nevertheless be useful to know beforehand what distractor compositions to use in the tasks, both the number of strategies used and the weighed number of strategies used were calculated. Also, as a variable, the number of different strategies used in an item was computed.

Item total correlations were computed, the point-biserial and biserial. These were entered as an index of item discrimination. The number of distractors selected by less than 5 percent of the test subjects were counted as a further indication of the attractiveness of distractors. This count was entered as a variable. A pre-estimate of the difficulty level of the word was used. Marzano and

Marzano's (1988) list indicates a level for a number of words at which the word is typically taught for foreign language students.

Item empirical curves were examined with TIAGRAPH software (Verhelst 1998) and the plots of distractors and items were examined. Based on the curve, the number distractors behaving as expected in different skill groups of the test population were counted and the count was used as a variable. Also, the behaviour of the correct choice was investigated this way, and it was entered as a binary variable according to whether the correct choice behaved as expected (1) or not (0).

3.6 Statistical procedures

Item parameters were computed using statistical techniques described in chapter 2.12. Then correlations were computed to examine the interaction of the variables in the study. Pearson correlations were computed for the second set of variables, but not for the nominal variables, as they have no meaning in that case. Inter-item correlations were computed for the performance data to assess the dimensionality of the data.

Principal component analysis was then used for the performance data to further investigate the dimensionality in the data. It was also hoped to find out what are the main sources in test structure contributing variance in the performance of the examinees.

Multi-dimensional scaling (MDS) as a descriptive method was used to investigate the test structure in a configuration with as many dimensions that were considered important based on the principal component analysis. The method also provides additional information about the dimensionality of the performance data. The main purpose of MDS was to group items based on their proximity, which would allow further examination of the test structure. Explanatory item / word features were further investigated with linear-regression, correlation and clustering techniques against the results of the MDS results.

The items were also clustered, and discriminant analysis was finally performed to investigate the relationship between clusters and and item / word features in order to see if these features can be used to explain the clustering of the items.

4. RESULTS AND DISCUSSION

4.1. Test quality

4.1.1 Item difficulty

Item difficulty was measured using two main sources. The item difficulty scale based on Rasch parameters is shown in appendix 7. The Pearson correlation between the difficulty figures is .775 (sig. .000) and they rank the items almost similarly (Pearson R=.908 sig. .000). The b-parameter values usually range between -3 and 3. It would therefore seem that two items - 10 and 13 - were easy for the examinees. The items test the knowledge of the words *form* and *hunting*. While these two items have ~~a~~ low percentages of failure, item 25 has the lowest of all items (less than one percent). The task assesses the knowledge of the word *reality*. However, the error involved in the estimation of the b-parameters was higher for the two easiest items (.304 and .447), while for ~~the~~ item 25 the error was only .252. This would seem to suggest that item 25 fitted better to the Rasch model, and it is therefore better suited for the present scaling purposes, while the other two demonstrate somewhat unexpected behaviour.

Because the p-value is a sample-dependent measure (see Baker 1997: 10), the Rasch-based item difficulty would seem more useful. Also, using the strict assumptions of the one-parameter model (i.e. equal discrimination indices), only ten items could be fitted to the model, but when discrimination was taken into account the discrimination of the items all items could be fitted to the model.

4.1.2 Item Disrcimination

Item discrimination was computed as the correlation between item and the total test (called the point-biserial correlation) using both the values computed by OPLM. A continuous variable is seen to underlie the dichotomous score in the biserial correlation (Baker 1997: 11). SPSS was used to calculate another index of discrimination, the biserial correlation. It is a Pearson correlation coefficient for a dichotomous (item) and a continuous variable (total test score). The last has been seen as a better measure of discrimination, although it varies somewhat systematically with the difficulty of the items and is affected by the degree ofheterogeneiety of the test sample (Baker 1997: 11). The discrimination estimates can be seen in appendix 8. The indexes have a correlation of .984 (sig .000) and values computed with OPLM have a slightly higher value, which can be expected (Baker 1997:11). The mean of point-biserial correlations is .338 and biserial correlations .278. The highest dicrimination power was observed for item 19 (r_{bis19} = .4843 and $r_{p\text{-}bis19}$ = .5480) and item 34 has the lowest discrimination power (r_{bis19} = .0532 and $r_{p\text{-}bis19}$ = .124). The distribution of discrimination is such that 25 items have a biserial correlation and 18 items have a point-biserial correlation higher than .300, and 38 items have biserial correlation and 40 items have a point-biserial correlation above .100. This indicates that the vocabulary items in general have a low discriminatory power. The results also show that, because of the variation in the discrimination of the items, it cannot be said that the discrimination is equal for all items. Therefore, the use of conditional maximum likelihood estimation should be used for computing the IRT difficulty parameters in order to have the most accurate results.

4.1.3 Test reliability

An important factor affecting the inferences made based on a test is the extent to which the items are free of error. In this particular test the items should focus on measuring vocabulary size, and anything else can be considered measurement error. Several ways have been developed within the framework of classical test theory for estimating test reliability. They are mainly based on repeated measurements, and assume that when an attribute (e.g. some specific point of vocabulary knowledge) is measured at least twice the amount of error included in the two or more occasions of measurement can be calculated from the variance in the test scores (Konttinen 1980: 25). The assumption in repeated measures reliability is that the measures are equal and made in circumstances that should lead to equal results in both situations, if error was not involved. The reliability estimate is, thus, an estimate of measurement consistency. When the testing situation is not based on repeated measures, it is possible to calculate an estimate for the lowest value of reliability in the test (Nummenmaa et al. 1996: 186). The estimate is called Cronbach's alpha reliability coefficient. This coefficient was calculated for the present test using the ALPHA subroutine in SPSS.

The task as a whole has a moderate reliability (alpha = .8068), and the elimination of any of the items would have no significant effect on this figure (alpha, if item deleted, is listed in appendix 5). The best item in the task, according to reliability, is item number 19, and the worst is item number 34. The subroutine also lists the reliability for the task, if any of the items were was removed from the set. A notable effect of the test length and reliability is that if items are added to ~~the~~ test reliability will improve, because the amount of variance attributed to error grows at a slower rate than the amount attributed to the ability being measured (Konttinen 1980: 40).

4.1.4 Model-data fit

When a model is used in connection with data, it is always necessary to know how well the model fits the data. With Rasch modelling it has been claimed that it is not easy to achieve a good fit for, for example, language test data, because of the assumption of unidimensionality. The fit of the model can be viewed from two points of view: global and the item level fit. The global fit is important in assessing the quality of the whole test. The item level fit indices tell about the behaviour of the individual items, and it can be used in identifying differentially functioning items (for example items that deviate from the unidimensionality assumption). OPLM reports the global fit with the use of R1c-test (Verhelst et al. 1995: 17). This test is based on a comparison of the expected and observed response patterns for the items in the test, and in a sense summarises item level information of fit. For the present test the approximation of the global fit was 192.242 with a significance of .6431. For a language test this indicates a good global fit.

gender

4.1.5 Differences between sexes and in age groups

Palmberg (1988: 207) notes that apart from the ability and the willingness to employ different comprehension strategies on a comprehension task, the difficulty of a task is also dependent on the test subjects' age, proficiency level and mother tongue. For this reason the distribution of abilities were tested in two different sub-populations. Variance in different groups was estimated using proportional success within the whole test population. First, the difference between the sexes was computed for individual items by calculating the difference of proportional success for the group consisting of women and men. The differences in the proportion correct for both groups as well as a plot of mean differences for items can be seen in appendix 6. The highest absolute difference was .150. This was observed for item 6 and the difference also indicates that the men were 15% more successful than the women on this item. The target word for the item is *estate*. One distractor was preferred over the other quite well for this item by women, and that was *väittää* (which can be translated in some contexts as *to state*). The high percentage (27.6) would seem to suggest that the form of the word has been a good distractor here for the examinees, who did not know the meaning of the word. The second best distractor for women was *estää* (which could be translated as *prevent*). The percentage was 9.6, and it is clear that the forms of the words are similar in appearance in the two languages, although there is no meaningful connection between the two words. The two words also represent different word classes. This would seem to indicate that the examinees who chose this distractor either did not have a response strategy to use for this item except for the similarity of the two word forms or that they found the other alternatives implausible in the situation. The percentages were similar in the men's group, but not as pronounced for the distractor *väittää* (17.0 and 7.1). It would seem that in this item the similarity of the form of the target word and the similar translation of the distractor word was a useful distractor.

Second, the difference in success on items was measured for three sub-populations composed of age ranges 18-28 (N=151), 29-41 (N=222) and 42-64 (N=102). The range of differences was greatest between the second and third groups (.610), then between first and last groups (.420) and last with first and second group (.361). The items with most differences between the groups were item 11 for groups one and two (p1-p2 = -.248), item 40 for groups two and three (p2-p3 = -.242) and item 19 for groups one and three (p1-p3 = -.319). The negative sign indicates that the items were easier for the latter sub-population, i.e. older examinees. Generally, success on the items was characterised so that 72.5% of the items were easier for the older sub-population between groups one and two, 55% between groups two and three, and 67.5% for groups one and three.

Item 11 had the most noticeable difference between age groups one and two. The target word for this item is *grease*. The preferred alternative after the correct choice was *hölmö* (which can be translated as *fool*). There does not seem to be a reason for this distractor except for plausibility. The second preferred distractor was ITAL possu (which can be translated as *piggy*). It is fairly similar to the Swedish word *gris* which means a piggy. This distractor qualifies as a useful distractor for age groups one and three, but not for the second age group.

The Most notable difference between groups two and three was observed for item 40. The target word of that item is *ward*, and it seems that this item is progressively more difficult in the different age groups. The preferred distractors after the correct choice were *holhooja* (*warden*) and *sarana* (*hinge*). It is again evident that the similarity between the target word and the English translation of the distractor is important in the first case, but for the second case the reason is not so obvious.

The difference between groups one and three was most notable with item 19, which had the target word *mean* as the stem. The distractor most of the people chose to take when responding to the item was *pääosa* (which refers to the lead actor of e.g. a play or a movie). It is also worth considering here the close relationship of mean to main which may provide the motivation for many people selecting this alternative. Many of the other alternatives were based on the more common sense of the target word, *average* (hence the distractor keskiaika, or *the Middle Ages*) or the similarity with the word *meat* (*paisti* can be translated to *roast meat*).

4.1.6 Representation of the semantic fields and word classes

Some of the features were not adequately represented in the test, or the items were placed into too many groups, and it seemed that these would contribute little to the study. To investigate adequacy of sampling, anti-image correlations were computed. The anti-image correlation for each variable against itself shows, if the sampling for each feature has been adequate. For this reason it is considered to be a measure of sampling adequacy (MSA). Abstractness of the word (R=.348), both the broad (R=.436) and precise (R=.441) classifications of semantic fields, broad semantic context area (R=.327) and frequency count for the Finnish word (R=.360) were not sufficiently represented in the data.

It was also noted that the test was not entirely free of sampling problems. The words were selected from a translation dictionary by using a simple random sampling technique (from a random point onwards every nth word was selected). This, however, seems to have resulted in an abundance of nouns in the sample (N=27) as opposed to other word classes (N=13).

4.2 Structure of the test and dimensionality

4.2.1 Correlations between variables

Examination of the correlation matrix for the word and item features tells about the association among variables in the data and, thus, about the structure of the test. It offers a possibility to check if the predictions from previous research hold also for this test. For this reason, the Pearson correlations were computed between the variables. The correlations are shown in appendix 9. Some of the correlations, however, are not the best possible for this task. For instance, the abstractness was entered as a dichotomous variable, and the correlation between a dichotomous and continuous variable tends show lower values compared to the actual correlation between the variables. The investigation, however, provides an overall impression of the test structure, and tells about the relationships of the variables in the study. The main interests in the investigation of correlations in this study is to assess the relationship of word features to item properties (item difficulty and discrimination). A coefficient was considered notable if it was above .300 and was statistically significant.

Word features that have a notable positive correlation with the classical item index include frequency ranking (R=.516**), number of homonyms (R=.463**), importance to the learner (R=.389*) and similarity of word forms (.376*). It was also noted that the number of contexts in which the word may appear has a notable negative correlation (R=-.338*), which indicates that words with several contexts tend to be easier. The b-parameter correlates with fewer features. It has the highest correlation with the similarity of word forms (R=.516**), and notable correlations also with number of homonyms (R=.400*) and frequency rank (R=.301*). A negative correlation was observed with distance from base word, indicating that base words were easier than derived and inflected word forms.

Correlations with the discriminability were investigated next. Two features had a high correlation with both indexes of dicriminability: similarity of word forms (for R_{bis} ,R= .537 and for $R_{p\text{-}bis}$,R= .580) and number of homonyms (for R_{bis} ,R= .445 and for $R_{p\text{-}bis}$,R= .488). Also, the number of contexts has a negative correlation with the point-biserial correlation (R=-.301). This would seem to indicate that there are discrimination problems with items which have a different appearance to Finnish words, have many homonyms, and tend to be met in several contexts.

Correlations with the variables that tell about how the items function are also important, because this tells about the link of word features to the quality of items. The absolute differences in different age groups and between the two sexes seem to indicate that there were differences between the last group and the two other groups. Only the differences between age groups 1 and 3 seem to be related to one of the word features, the number of homonyms (R=.461**). On item level, the group differences are related to reliability, discrimination and the classical difficulty index for items. The differences in item difficulty can be traced to differences between group 3 (the oldest examinees) and the other two groups (for groups 1 and 3 R=.538** and for groups 2 and 3 R=.564**). This means that the differences are greatest with difficult items, and the oldest age group seems to be in a different position to the other groups. The differences also correlate with the number of unattractive distractors (R=-.430** and R=-.419**), which indicates that differences are greatest with items that have less unattractive distractors. This would seem to suggest clear differences in performance of the group of oldest examinees in the test.

The absolute differences between the women and men seem to be related to importance and abstractness. Importance correlates positively (R=.345*) and abstractness negatively (R=-.435**). This suggests that words which raters though were important for the learners tended to cause differences between the sexes. Also the abstract words tended to cause less differences between the sexes, while concrete words were those where there was greater difference. Because the absolute differences were examined here, it is possible to tell where the differences can be observed, but this does not tell if the men or women had an advantage in responding to the item.

Number of unattractive distractors correlates positively with number of contexts (R=.489**) and negatively with importance (-.465**), similarity (-.438**), frequency rank (R=-.417**) and number of homonyms (R=-.413**). This indicates that words which can be met in several contexts, and which the raters considered less important for the learners, and which are less frequent and have less honomyms tend to appear in items which contain more unattractive distractors. The number of distractors in an item that work correlates positively with the number of homonyms (R=.354**) and importance (R=.334*), and negatively with the number of contexts (R=-.436**). This indicates that guessing can be associated with those words for which there are more homonyms and were considered important for the learner and can be met in few contexts. It was also noted that items for which the TIAGRAPH showed problematic results for the correct choice, the target words tended to be dissimilar in the two languages (R=.461**), infrequent (for LOB R=-.474** and BNC R=-.406**), long (R=-.390**), more imageable (R=-.342*), less important (R=.323*), and had a lower number of homonyms (R=.322*). These associations seem to suggest that it may be possible to have an effect on the quality of items with reference to four variables: number of contexts, similarity, importance and number of homonyms. It may be that it is more difficult to make good items for these kinds of words, and these cases should be constructed with care. It was also noted that similarity (R=.514**) and number of homonyms (R=.456) have a positive relationship with the reliability of the test. This suggests that when the words in the test had a similar appearance and had a low number of homonyms, they tended to enhance the reliability of the test.

The first classification of distractors correlates positively with the estimated level on which the word is taught first (R=.313*), and frequency (LOB R=-.301 and BNC R=-.314*). If the level estimate and frequency are considered as a pre-estimation of the level of the word (and, hence, at least partly of the item) this seems to indicate that the number of different kinds of distractors of the first classification are somewhat related to the presumed difficulty of the word. The second classification correlates with abstractness (R=.345). This suggests that from the point of view of psycholinguistic error sources, items that had an abstract word as the target word tended to contain different kinds of distractors.

4.2.2 Comparison of frequency counts

In order to assess the relationship of the two English and one Finnish frequency word lists Pearson correlations were computed for the item words used in this study. The correlations show that the BNC and LOB word lists are similar to each other (R=. 995, sig.=.000). The Finnish word list does not correlate with either the BNC word list (R=.001) or LOB (-.010).

It, however, seems that an index based on the rank order of the words in the frequency word list is more useful for assessment purposes. This index correlates positively with the classical item difficulty index (R=.516) and the b-parameter (R=.301), while the correlations for the raw frequency counts are below .300. This is most likely due to the fact that the raw frequency counts may emphasise the differences between less frequent and rare word.[5]

4.2.3 Dimensionality

In order to establish a measurement scale, it is useful to see how many dimensions underlie it. A dimension is thought to be a feature of measurement, and usually measurement is limited to one dimension. This means in practice that the test measures one ability only and is affected very little by other abilities. However, it is not possible to measure language abilities totally in isolation of other skills. For example some feature of the vocabulary test items may also measure knowledge of grammar, and hence the vocabulary dimension may be affected by the examinees mastery of that grammar knowledge that is also required to complete the item. The interaction from other skills is treated as a source of measurement error and hence unreliability in unidimensional tests.

For classical test theory, the measurement dimension is represented by the true score and it describes a single ability that the items are tapping. In this study the dimension the items are tapping should be the size of vocabulary. For Rasch analysis the dimension is the ability-difficulty continuum.

Rasch analysis based on the one parameter model should be used for data that is unidimensional, i.e. there is only one dimension being measured. This has been a fact of considerable controversy recently, but currently the consensus is that if the strict assumptions of Rasch modelling can be applied to the data set, then Rasch modelling is useful. However, it is also necessary to establish dimensionality in order to be certain that the inferences made regarding the test are valid.

There are several methods available for assessing dimensionality of research data (see e.g. Henning 1988: 84 and Hattie, J. 1985: 141), mainly because of the fact that it is easier to make inferences based on a unidimensional measurement instrument. It should also be stressed that none of the methods can be used to show for sure that a test measures only one dimension. For this reason, it is common to apply several methods. For this study, inter-item correlations, analysis of eigenvalues in principal component analysis and stress levels of multi-dimensional scaling were used.

4.2.4 Inter-item correlations

The dimensionality of the items were studied first by computing inter-item correlations. Inter-item correlations are used to estimate roughly the dimensionality in the performance data. Since the data used in this study is dichotomous, it was more useful to compute tetrachoric correlations instead of ~~the~~ Pearson correlations. Kendall (1967: 304) describes the use of the tetrachoric correlation for examining the asscociation between dichotomous measures that are theoretically thought to measure a continuous concept; in other words, a concept on which a dichotomy has been imposed. Also, Carroll (1961: 349) shows that Pearson correlations may mask the distribution of the. He (1961:357) suggests computing tetrachoric correlations to control this behaviour.

Tetrachoric correlations were computed using OPCOR (Verhelst 1998) software. In the computation of the correlation coefficient two problems were observed for some items. When the p-value of one item is zero or one, the correlation cannot be computed, because the algorithm does not converge. The second problem observed was with item matrices which include too many zeroes. In this case the correlation does not exist. (Verhelst 1998: 2). Both of these cases resulted in missing values for the correlation.

Connected items	1 2 4 5 7 9 11 12 14 15 16 17 18 19 20 21 22 24 26 27 28 29 30 31 34 36 37 38 39 40
Unconnected items	3 6 8 10 13 23 25 32 33 35

TABLE 4.1. Connected and unconnected items according to tetrachoric correlations.

The program makes a suggestion about which items are connected. Ten items in the present set were disqualified in the process, and these are shown in table 4.1. The examination of inter-item correlation proceeded with using the existing correlations.

Once the tetrachoric correlations have been computed, the estimation of dimensionality can be based on the distribution of the inter-item correlations. Piedmont and Hyland (1993: 370) suggest that a unidimensional scale would be characterised by items that correlate equally well with each other. The correlations should, in that case, have a mean of 0.3 with a normal distribution of the correlation coefficients. The more dimensions a scale has the more correlations will be zero. In reality, however, the results have to be interpreted in relation to human performance, as no performance is ever perfect, and therefore these rules should be applied accordingly. The distribution is therefore allowed to be slightly skewed positively, i.e. tilted backwards towards zero.

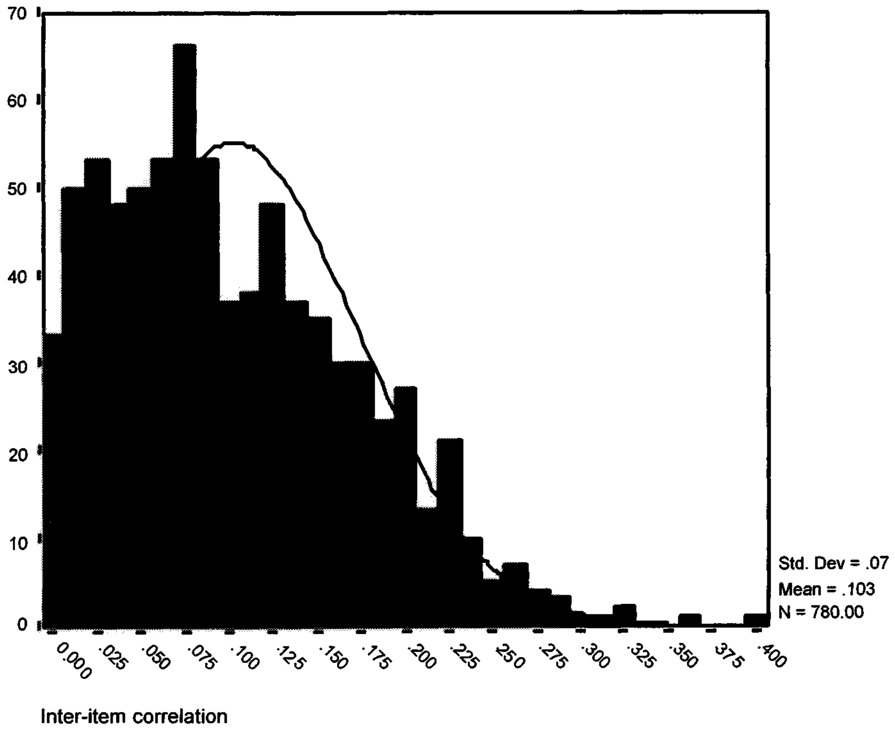


FIGURE 4.1. Distribution of inter-item correlations with comparison to normal distribution displayed.

Figure 4.1 shows the distribution of inter-item correlations. Based on the performance data, the underlying number of dimensions would seem to be one. The mean of the correlations is .24 and skewness is .591. There is therefore reason to believe that the underlying data structure is unidimensional. This was further analysed with the help of principal component analysis.

4.2.5 Principal component analysis

Another way to examine the dimensionality of the data is to perform principal component analysis. The analysis determines the underlying structure of the data by examining the correlation or covariance structures of the variables that were measured. In other words the method attempts to identify parts of the variance in the measures that can be said to come from the same source. Harris (1975: 163) says PC can be used to investigate uncorrelated contributions that make up the data. Since it is desirable for a test to be unidimensional, ideally a single PC would indicate a truly unidimensional test. But since human behaviour is inherently affected by other factors in addition to the one being measured, it is rarely possible to find only one PC. However, the discovery of one dominant PC and other components with low loadings, is usually accepted as evidence for unidimensionality (Hambleton 1985: 157). Choi and Bachman (1992: 59) used principal component analysis for several vocabulary and reading tests and obtained a first component explaining between 31.5 to 40.6 per cent of variance, and used an acceptability level of 20 percent of variance for the first component.

The analysis (was) is performed on the correlation-covariance matrix of the data. In this instance, it is important to use the correct type of correlation. This would naturally be the tetrachoric correlation used earlier in the calculation of inter-item correlations. However, the PC extraction process involves inversion of the correlation-covariance matrix, in which it is possible for the determinant of the correlation matrix to become undefined positively. While Pearson correlations are guaranteed to have a positive determinant, this is not necessarily the case with tetrachoric correlations. It is possible for the determinant to become negative or zero. The last instance is also the case with the current data even when missing correlations are excluded from analysis. For this reason, the Pearson correlation was used, even though it does not provide as accurate results.

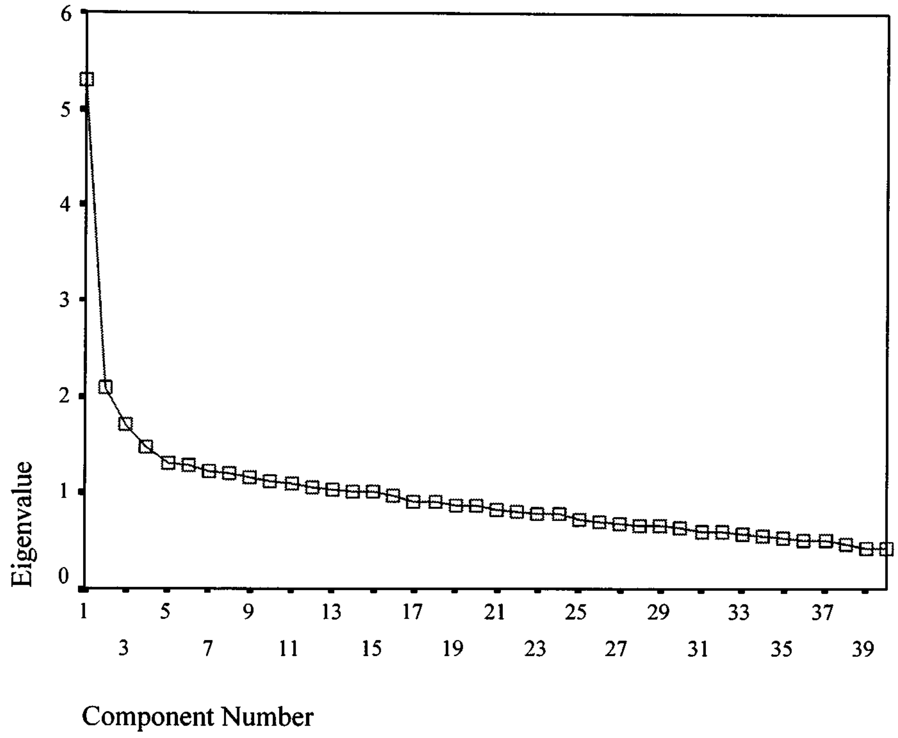


FIGURE 4.2. Scree plot for values of eigenvalues.

Factor	Eigenvalue	% of variance
1	5.29309	13.2
2	2.09325	5.2
3	1.70694	4.3
4	1.47506	3.7
5	1.30539	3.3
6	1.28422	3.2

TABLE 4.2 Eigenvalues and percentage of variance for the first six factors of performance data.

Table 4.2 shows that one major factor was found for the performance data that explained 13.2 percent of variance. The percentage is not very high, but it is clearly higher than the amount of variance explained by the following factors. The difference can be visually seen in the scree plot in figure 4.2.

The number of explanatory PCs is usually interpreted in terms of the amount of variance they explain. Few rules exist in determining the number of significant PCs. One way is to use the scree plot as an aid. The number of factors can be visually estimated by looking at a point in the cline, where it starts to level off (Kim and Mueller 1978: 44). For the performance data this happens at the point of the fifth factor. On the basis of this analysis the test would seem to be unidimensional, although the number of low loadings on the other components suggests that also other aspects have an influence in the test.

4.2.6 Stress levels of multi-dimensional scaling

A further method to investigate dimensionality is to use the stress-levels of multi-dimensional scaling. In the analysis the number of dimensions is controlled by the analyst, and the analysis calculates a solution on the basis of the performance data that shows the location of items in a n-dimensional space. A stress value is calculated to show how well the selected number of dimensions account for the structure obtained and that the structure fits the data well. Low stress values mean that the dimensions correspond to the number of dimensions well. However, it is also possible that meaningless dimensions are included in the model when too much reliance is put on the stress value alone. These dimensions contain information about noise, i.e. variation that is not interesting for the subject being measured. These dimensions may have utility, however, which will be shown later. For this reason, it is necessary to compute the stress levels for solutions with different numbers of dimensions, and see what is the most useful number of dimensions.

Kruskall and Wish (1973: 54) recommend two methods for assessing the number of dimensionality in the data. One is based on observing the stress level in the data, while the other relies on experience. Both methods were used in this study, although the emphasis is on the first method when the number of dimensions is examined. It involves using Monte Carlo simulation of random data with known number of dimensions. The simulation data can be constructed so that the dimensionality of the data is known, but also the amount of measurement error can be included. The stress levels observed in the simulation configurations can then be compared to those in the actual data to gain support for the most likely dimensionality of the real data.

Since multidimensional scaling examines the similarity in the data, it was noted that some examinees may employ different response strategies, because of their language background. It was considered that this would have a more profound effect for those examinees that took the Swedish test, since they took items that were translated from Finnish. The translated items were considered to be essentially different in comparison to the Finnish items. Therefore they were excluded from the further examination of dimensionality. When the stress level was examined excluding the Swedish-speaking examinees, stress levels were slightly lower, but the results show that they contain a high level of error. Thus, the stress levels were calculated for up to a five dimensional configuration for the performance data. Kruskal stress level for the five cases improved slightly, when dimensions were added (.23397 in one dimension, .16066 in two dimensions, .12424 in three dimensions, .09407 in four dimensions, and .07730 in five dimensions).

Figure 4.3 shows the stress levels related to the present data in comparison to Monte Carlo simulations. The simulated data have been constructed with a known level of error, which has been marked in the figure. The line marked with crosses is a threshold value for infinite error, and the line marked with stars includes error on a level of .25. It can be noted that the solid line representing the dimensionality in the actual data is notably flat, and is very close, but below error level .25. When the number of dimensions is increased, the amount of error increases. Kruskall (1973: 54) states that in this case the configuration with one dimension is likely to apply for the data, but it is not an extremely useful case for making inferences. Optimal one dimensional solution would have a stress level of .15 or below with a rather flat line of lessening stress levels. The plot of the stress level should also exhibit a marked elbow at the point of real dimensionality, i.e. for one dimensional solutions the plot would show a line with no dramatic turn in the line. This can be observed for the present data and it suggests a unidimensional configuration to be most likely. The stress level of the unidimensional solution is rather high, and it seems to increase when the number of dimensions is increased. The presence of high level of error in the stress levels seem to indicate that may be evidence of guessing and individual solution strategies of the examinees have an effect on the estimation of abilities, but support earlier findings of a unidimensional test.

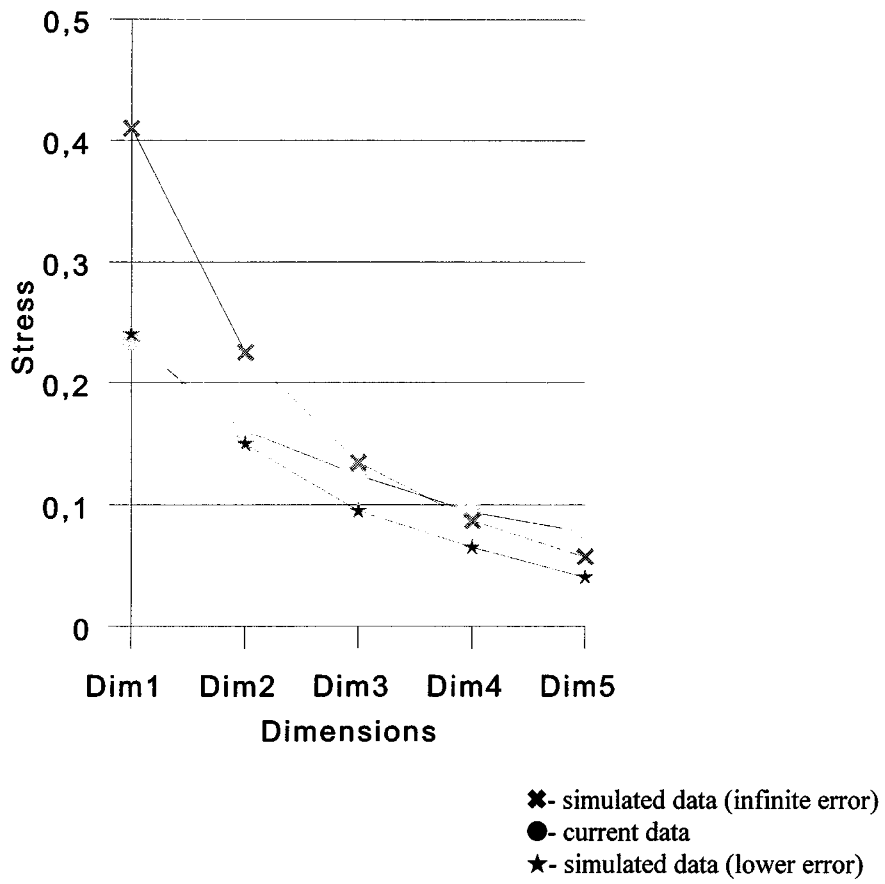


FIGURE 4.3. Stress values in the actual and simulated data.

The second method for assessing is more subjective. Stress level tells only how well the configuration obtained matches the original data set, and for this reason it cannot be accepted alone as evidence of dimensionality. Useful interpretation of the configuration may, however, necessitate the use of more than one dimension, even though a reasonable interpretation cannot be found for all of the dimensions. It seems that based on the principal component analysis that at least two dimensions are useful for the interpretation of the results. Interpretation is not based solely on the statistical assessment of the number of dimensions, but takes also into account the stability, ease of use and interpretability of the configuration (Kruskal 1973: 56). The interpretation, however, has more utility to other purposes of this study, and it will be attempted below, after the structure of the data is examined more closely.

4.2.7 Item clustering

The distances between items estimated with multi-dimensional scaling can be used to arrange items into groups according to their distances. The proximity values provided by the PROXIMITIES subroutine of SPSS were used to form clusters of the items. Kruskal (1973: 88) suggests that difference values below .300 should be used as a threshold value for including items in a group, beyond that the distances are too far away for meaningful connections to be warranted. Groups were formed in this way for all five different groups of MDS configurations. Of most interest were the one, two and three dimensional configurations that can be plotted, and the figures show one dense group that contains several items, and the configurations up to three dimensions contain most of variance in the stress values. The groups can be seen towards the right in the histogram in figure 4.4.a., and towards the centre left in the figure 4.4.b. The other groups in these configurations have considerably fewer items in them. The groups formed for one and two dimensional configurations can be seen in table 4.3. The groups formed for the three, four and five dimensional configurations all contain one major group and most of the rest of the items are unconnected with each other, so that in the three dimensional

Dimensions	Group	Items in group
1	1	2,5,6,12,13,18,19,23,27,28,29,32,33,34,38,39,40
	2	3,8,9,14,17,20,22,24,30
	3	1,7
	4	11,16,21,25,31,37
	5	4,10,35
	6	26,36
	7	15
2	1	2,3,4,7,8,9,10,13,14,23,25,26,33,34,35,38,39
	2	11,17,21,24,29,30,36
	3	1,6,15,19,37 (31)
	4	12,40
	5	16,20,27
	6	22,32
	7	5
	8	18
	9	28

TABLE 4.3. The groups formed of test items based on similarity of response patterns.

configuration 26 groups could be identified, 30 groups in the four dimensional, and 31 groups in the five dimensional configuration. After this grouping of items, the interest was to see which of the word and item word features would be useful in describing the differences between the clusters.

4.3. Relationship between content-based item descriptors and empirically-based item characteristics

4.3.1 Correlation with item difficulty

In order to see what were the possible features that describe progress from lower to higher ability, the correlations among item difficulty and other variables were examined. Table 4.4 shows correlations for the word and item features against Rasch item difficulty. The first column shows the correlation coefficients for the features. The

Word / item feature	Coefficient	Sig.(2-tailed)
abs_con*	-.273	.088
wordclas	N/A	
freq BNC	-.118	.017
freq LOB	-.125	.490
freq FIN**	-.073	.679
freq rank	.301	.075
context	-.274	.123
no_hom	**.400**	.011
imageab	.003	.983
no_let	-.283	.140
distnc	**.516**	.001
imprtnc	.233	.148
base_dis	-.307	.054
no_sfLOB**	-.190	.282
#CLAS1_1 NULL	.105	.520
#CLAS1_2 FORM	.088	.590
#CLAS1_3 MEANING	-.156	.336
no_CLAS1	**.400**	.010
#CLAS2_1 Transfer	-.266	.097
#CLAS2_2 Intralingual	.078	.632
#CLAS2_3 Unique	.104	.521
no_CLAS2	-.071	.663
*Semantic field (precise)**	N/A	
*Semantic field (broad)**	N/A	
alpha	**-.546**	.000
age1_2	**.404**	.010
age1_3	**.514**	.001
age2_3	**.445**	.004
gen_diff	-.157	.332
inv_p	**.775**	.000
itTotSPS	**.603**	.000
itTotOPL	**.707**	.000
OPLM_SE	**-.920**	.000
#WCLAS1_1 NULL	**.541**	.000
#WCLAS1_2 FORM	**.511**	.001
#WCLAS1_3 MEANING	.142	.384
#WCLAS2_1 Transfer	.226	.161
#WCLAS2_2 Intralingual	.340	.032
#WCLAS2_3 Unique	**.411**	.008
itemattr	**-.725**	.000
tiagrafd	**.637**	.000
tiagrafc	**.455**	.004

TABLE 4.4. Correlations between item difficulty and word/item features.

second column shows the significance of the coefficients.

The results show that the features that seem to work best with item difficulty relate to item features rather than the qualities of the target words. They include indicators of error such as the alpha value if an item was removed from the test and Rasch standard error for the b-parameter. The former shows that item reliability seems to be closely linked to item difficulty, i.e. less difficult items do not seem not to be as reliable as more difficult items. Classical discriminability indexes have a similar relationship: difficult items discriminate better as can be seen for item-total correlations from SPSS and OPLM.

The weighted item properties indicate that correlations for different item construction strategies are better for item difficulty. Distractors based on form or on no apparent reason work better than those based on meaning. Also, intralingual errors seem to be more attractive distractors than those that rely on transfer or unique errors.

The correlations were not high for the frequency counts, which suggests that the word lists as such are not very useful for predicting item difficulty. It was, however, more useful to use the rank number of the word as an indicator of difficulty, because this had a better correlation with item difficulty. the Number of contexts also seems to have a slight negative correlation with item difficulty, close to that of the frequency indexes.

From the table it can be seen that in constructing items, it is useful to take into account the frequency rank of the target word, number of its homonyms and the similarity of the word form to a mother tongue or third language word. It would also seem that items become more difficult when the number of different kinds of classification 1 type of distractors is increased.

4.3.2 Principal component structure

The structure of the performance data can be examined with reference to the word and item features. The aim is to investigate which of the features can be associated with performance on the items. Harris (1975: 158) makes an analogue from principal component analysis to an internal discriminant analysis, as the PCs represent a linear

FEATURE	PC1	PC2	PC3	PC4	PC5
abs_con	.109	.064	-.354	.010	-.126
base_dis	.109	.001	-.334	.203	-.081
basic	.105	.113	.118	-.089	**-.339**
context	-.058	-.029	-.133	-.188	-.275
distnc	.039	.223	**.329**	.243	.244
freq BNC	-.141	.530	-.110	-.045	-.037
freq FIN	.042	-.128	-.073	-.090	-.051
freq LOB	-.158	.521	-.148	-.030	.014
freq Rank	-.011	-.142	-.184	.321	**.575**
imageab	**-.335**	.154	-.273	-.074	.074
imprtnc	.006	-.104	.177	.272	**.340**
level	.077	-.175	.344	-.049	.115
no_ hom	.037	.015	.026	**.323**	-.137
no_ let	-.235	-.067	-.113	-.123	.166
It_cl2_1	.063	.097	.076	-.087	-.041
It_cl2_2	-.061	.130	-.070	-.097	-.110
It_cl2_3	.019	-.200	.019	.160	.142
it_cl1_1	.159	-.196	**.345**	-.116	.019
it_cl1_2	-.070	.104	-.045	-.126	-.239
it_cl1_3	-.083	.078	**-.314**	.285	.284
no_str_1	.129	.014	-.279	.042	-.008
no_str_2	.094	-.042	-.042	.048	.225
age1_2	-.107	-.033	-.106	.000	.142
age1_3	-.151	-.135	-.055	.076	.294
age2_3	-.076	-.044	.121	.079	.302
alpha	.176	.150	.055	-.046	.013
gendiff (absolute)	-.117	.118	.152	.142	-.047
inv_p	-.125	-.098	-.051	.243	**.501**
itTot_SPS	.188	.130	.110	-.040	.004
itemattr	.129	-.064	-.014	-.298	-.241
itTot_OPL	.153	.138	.093	.009	.080
OPLMDiff	-.064	.146	.130	.148	.297
OPLM SE	.021	-.274	-.188	-.072	-.116
SUMWC1_1	-.191	.133	-.104	.253	.336
SUMWC1_2	-.045	-.251	.128	.111	.333
SUMWC1_3	.063	-.007	-.199	.070	.102
SUMWC2_1	.029	.026	.033	.241	.328
SUMWC2_2	-.088	-.253	-.022	.082	.400
SUMWC2_3	-.076	-.038	-.050	.074	.032
tiagrafc	.243	.071	.308	.238	-.251
tiagrafd	-.026	.244	.009	.079	-.044

TABLE 4.4. Correlation of item / word features and principal components.

combination of the original variables, which maximally discriminate between the subjects. In this way it is possible to find out the degree to which the features affect responses on the items. Thus, Woods (1983: 50) states that principal component analysis is a descriptive method, and it is not based on a model as is the case with factor analysis. The usefulness of PCA is in its ability to express the variance on items in a condensed form, and also how the response patterns differ between test subjects, when the research question is explorative as is the question here (Tabachnick 1989: 28).

Principal components can be interpreted with the help of multiple regression in order to explain the structure of the PCs loadings obtained for the original variables (Marascuilo: 239). In the present study the rotated PC loadings were compared with the features of the words and items. For this reason, the principal component was rotated using varimax rotation, which yields orthogonal, i.e. uncorrelated components. Table 4.5 shows the correlations between the components and the word and item feature variables.

The importance of each component was already looked at earlier, when the eigenvalues of the factors were looked at. The first component is responsible for most of the variance, and is thus the most important component underlying the data. The highest positive correlations can be found for the difference between sexes (.288) and the correctly modelled correct choice (.243). The highest significant correlation is with imageability (-.355*). Other high and negative correlations include importance (-.241) and number of letters (-.235). It seems difficult to name this variable except in terms of these variables, which indicate that when the importance and the imageability of the words being tested decreases, the differences between sexes become more evident, and these also affect the modelling properties of the correct alternative.

The second component seems more closely related to the frequency counts of the words (BNC R=.530** and LOB R=.521**) as well as the similarity of the English word to a Finnish word (R=.223). It has also negative correlations on some of the weighted distractor strategies and the magnitude of error of estimating Rasch parameters.

The third component correlates positively with the number of form-based distractors used in an item, level on which the word is usually introduced for the learner and similarity of word forms in English and Finnish. It has negative correlations with the abstractness of the word, number of affixes, and number of unclear distractors.

The fourth component is dominated with number of homonyms and with word frequency. The last component can be associated with word frequency, inverse p-value, importance of the word, four weighted distractor error classes. The fact that the correlations on this component are substantial means that they can be named with most precision. However, the fact that the last three components account for a small amount of variation means that they have little practical utility.

4.3.3 Multi-dimensional scaling

It was already noted that multi-dimensional scaling is used to assess distances between variables based on the similarity of the response profiles to items. The method has been used in studies of similarity of responses to psychological stimuli such as how colors are perceived. The method locates variables on an n-dimensional space according to the similarity of data patterns. In this space similar response patterns are close to each other while different patterns are located apart, so that the more extreme the difference between two variables is the farther away they lie from each other.

The main interest in examining response patterns in this way is to identify persons who are using similar response strategies. It can be assumed that people who employ similar strategies will have similar response patterns and they will be located close to each other in the dimensional space. The investigation can also proceed in another way, when the interest is on mapping the location of items in relation to each other. Then the proximities of the items show how close the items are in terms of the response patterns. It can be assumed that items, which have a similar response pattern in the performance data, require the use of similar response strategies. If some items are close to each other, it is possible to investigate if some of the descriptive features of items are responsible for this similarity. This basic idea has also been used successfully in connection with lattice theory (Young 1998). The aim of using the method in this study was to see which of the items are more alike according to the performance data, and what features of the items would explain the similarity.

This actual method for mapping the items in the space of n-dimensions is similar to factor analysis. Where factor analysis attempts to reduce the number of measurement dimensions to as few as possible, MDS starts from one dimension and the number of dimensions can then be increased to as many as the analyst wishes in order to see which number of dimensions provides a meaningful interpretation of the data structure (Konttinen, 1997: 291). The method used for this study was the Euclidian distance model, in other words concrete measurable distances. Other models are also possible, but this one is usually preferred because of ease of interpretation. Kruskal and Wish (1973: 45) state that the selection of the model is not as important as the interpretability of the results.

It was noted in the investigation of stress levels that one dimensional solution would explain a considerable amount of variance in test responses. This, however, refers to the measurement quality of the test. One major dimension can be claimed to underlie the data, but at the same time it was noted that five components of the PCA data had significance in explaining the variance in data, and up to three dimensions were identified with a known level of error. Useful interpretation of the configuration may necessitate the use of more than one dimension in describing the distances, even though a reasonable interpretation cannot be found for all of the dimensions. The interpretation is not based solely on the statistical assessment of the number of dimensions, but should also take into account the stability, ease of use and interpretability of the configuration (Kruskal 1973: 56). Because the principal component structure suggests that five components explain most of the variance, the interpretation of configurations was attempted for the five dimensional configuration.

For one and two dimensional configurations it is possible to plot the coordinate values and try to interpret the meaning of the dimensions by looking at the groups of items. This is, however, difficult if the number of potential explanatory features is high, as is the case in this study. In this case, Kruskal and Wish (1978: 36) recommend the use of linear regression as an aid. Regression in this way attempts to find an optimal line that, when fitted in the plot captures most of the locations of items in the configuration. For the linear regression process, the item and word features were treated as the dependent variables and they were regressed over the two sets of coordinates.

Tables 4.6a and b show the multiple correlations of word and item features against the five dimensional configuration. Next to the multiple correlations are the regression coefficients for the features on each dimension. Since the coordinate axes are independent of each other, a straight line that captures most of the variation for the dimensions is the most likely explanation for that dimension. It is applied by looking up the most significant multiple correlations (columns two and three) and then looking at the strength of the regression coefficient for each dimension (Kruskall 1973: 37-40).

The first dimension in all configurations could be identified to a fair degree of accuracy, while the definition of the second dimension was less clear. Multiple correlations are high for classical item difficulty index, number of distractors chosen by less than 5% of examinees, Rasch difficulty parameter, Collins and Cobuild frequency count and number of synonyms. While the difficulty values and the frequency count all have a positive regression coefficient, the number of distractors has a negative value. This expresses the tendency of items becoming more difficult while the number of unattractive distractors decreases. These correlations are significant at the .000 level. Less significant correlations are observed for rank order of items based on the Rasch difficulty parameter, number of distractors that are not based on any observable strategy, and importance for the language learner. The first dimension can therefore be called the *difficulty dimension.* The second dimension has the highest correlation with the number of synonyms, nearly as high as for dimension one, and negative correlations for the item difficulty parameters.

Correlations are low for the other dimensions, and Kruskall (1973: 37) does not recommend putting too much weight on forcing an interpretation on such a dimension with low correlations especially in a case where the two dimensional solution seems to contain a high level of error. It seems more likely that the high coefficient on dimension one for the number of synonyms has an effect on the mastery of items. If a word form has

Feature	One dimensional configuration		Two dimensional configuration		
	Multiple correlation	Regression weights	Multiple correlation	Regression weights	
		Dimension 1		Dimension 1	Dimension 2
unattr	.746	.746	.751	-.742	.064
no_cl_1	.293	-.293	.300	.288	-.063
no_cl_2	.003	.003	.308	.249	.207
it_cl1_2 (form)	.023	-.023	.027	.025	.014
it_cl1_3 (meaning)	.341	.341	.342	-.341	.004
it_cl1_1 (null)	.410	-.410	.411	.409	-.019
wit_cl1_2 (form)	.125	.125	.751	.752	.007
wit_cl1_3 (meaning)	.332	.322	.553	.525	-.131
wit_cl1_1 (null)	.071	-.071	.139	.131	-.035
it_cl2_1 (transfer)	.068	.068	.193	.103	.154
it_cl2_2 (intraling.)	.023	-.023	.219	-.140	.156
it_cl2_3 (unique)	.070	.070	.358	.215	.267
wit_cl2_1 (transfer)	.092	.092	.429	.404	.189
wit_cl2_2 (intraling.)	.354	.354	.451	.452	.009
wit_cl2_3 (unique)	.125	.125	.430	.432	.022
tiagrafd	.347	-.347	.380	.378	.090
inv_p	.916	.916	.932	.916	-.102
gendiff	.135	-.135	.131	.129	.037
tagrafc	.268	-.268	.312	.295	.133
b-parameter	.679	-.679	.690	.338	.008
SE for b-par.	0.079	-.079	.449	.690	-.158

TABLE 4.6a. Multiple correlation and regression coefficients for clusters and item features.

Feature	One dimensional configuration		Two dimensional configuration		
	Multiple correlation	Regression weights	Multiple correlation	Regression weights	
		Dimension 1		Dimension 1	Dimension 2
number of contexts	.277	.277	.302	-.241	.167
number of letters	.262	.262	.267	-.267	-.050
number of homonyms	.418	-.418	.598	.488	.395
abstractness	.062	.062	.188	.030	.182
importance	.387	-.387	.374	.366	-.148
similarity	.328	-.328	.321	.321	.002
imageability	.106	-.106	.111	.110	-.011
distance to base word	.231	.231	.251	-.217	-.108
Frequency rank	.567	-.567	.640	.496	-.360
Frequency (LOB)	.164	.164	.169	-.158	.050
Frequency (BNC)	.175	.175	.186	-.164	.073
Frequency (FIN)	.139	.139	.136	-.136	-.002

TABLE 4.6b. Multiple correlation and regression coefficients for clusters and word features.

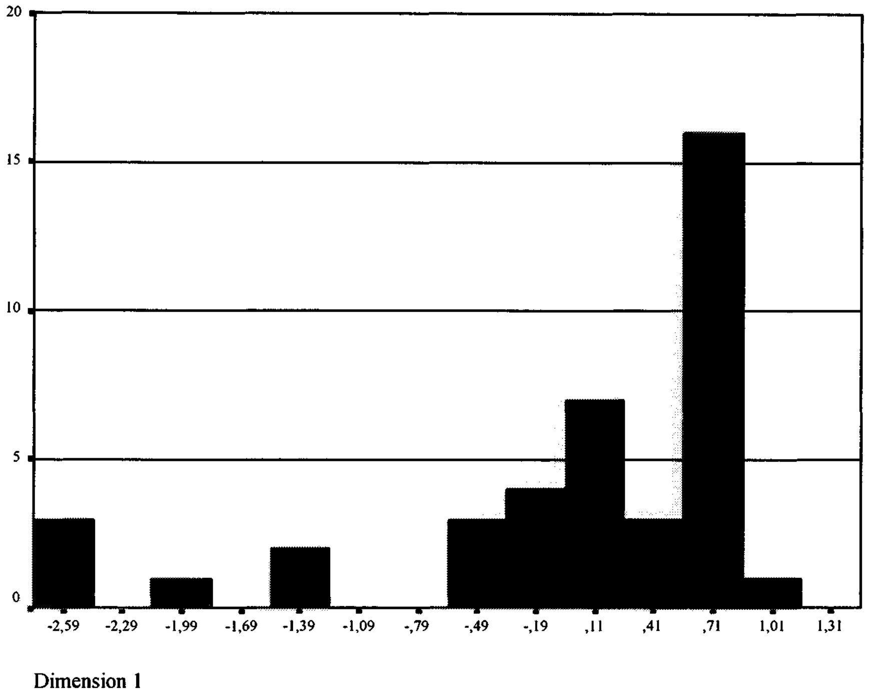


FIGURE 4.4.a. Histogram of item clusters in one dimensional space.

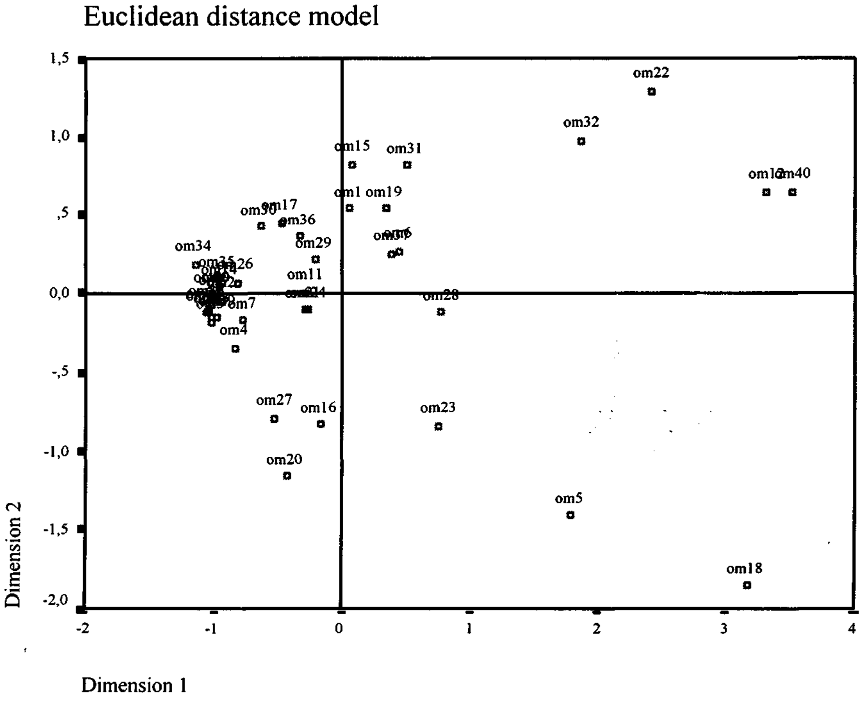


FIGURE 4.4.b. Plot of item clusters in two dimensional space.

several senses, it is possible that the learner does not know the one being supplied as the correct option.

An alternative attempt can be made at interpreting the configuration. For purposes of the subsequent analysis this method was used for all the five different configurations. The method is based on the clusters that items form, and the dimensional plot of the items can be used as a help to identify clusters. Figures 4.4a and 4.4b show the item plots for the one and two dimensional configurations, but it is not possible to plot a configuration with more than three dimensions. Histogram is used for the one dimensional space, because it is easier to visualise the item clusters in this case. Both plots have a high density of items at one point of the multi-dimensional space. Dimension one in the two dimensional cluster, which was noted to have a relationship with item difficulty, indicates that the large cluster contains several easy items. Items seem to scatter in an increasingly diffuse pattern towards the right. Items in the second cluster are also further apart on the second dimension. An examination of the items reveals that they seem to be alike, i.e. they have been designed using similar distractor strategies. However, the word features seem to indicate that items with positive values refer to words that are abstract and have a generic reference, while words with negative values are concrete and refer to specific concepts. This observation is based on a small number of items present on the right hand side of the figure.

4.3.4 Discriminant analysis

Discriminant analysis is used to investigate the accuracy of classification (Klecka 1973: 42, Tabachnick 1989: 507). The analysis investigates the effect of a group of independent variables on discriminating between specified groups. It shows if the information from the independent variables can be used to assign the cases into the same groups that are specified. If the prediction is found to be statistically significant, the variables can be used to predict results with future samples of items and examinees.

The groups were formed, as described above, from the three dimensional configuration of the multi-dimensional scaling results. This is because this number of dimensions was observed to explain 22.7 percent of variance in the principal component analysis, and also because the investigation of the stress levels of the multi-dimensional scaling results indicated that three dimensions are all below the error level for noise. The variables representing word and item features were used as the independent variables. This setting, thus, tells us to what extent the word and item features account for the similarity of a group of items classified according to the profiles of item responses. It was also considered useful to assess which feature sets are best for use in classification, and for this reason the word and item feature variables were entered in four different sets in separate analyses. This was done because in different phases of the test construction the test constructor has a different amount of information available, and it is important to know to what extent that information can be relied on. First, all of the features were used to assess the usefulness of the variables as a whole. Second, the word features were entered alone in order to see the usefulness of word features. In the third step, all variables which could be used before the test was administered, were included in the analysis. The aim was to see to what extent future test construction could benefit from knowledge of the items features. In the last step, only item features were used. Appendix 10 shows the variables used in the different steps of the analysis.

Tabachnick and Fidell (1989: 510) claim that 95% accuracy is sufficient for classification purposes, and in this case it is not necessary to meet the restrictions concerning the shape of distribution of the independent variables. The classification results for the different configurations in the different variable combinations can be seen in table 4.7. It can be seen that a notable loss of prediction accuracy for all features occurs at three dimensions. It was also observed with regards to the stress value that the amount of error included in the measurement also becomes high at this point. Subsets of word features contain useful prediction value in the three dimensions, and the prediction rate drops slightly for the four and five dimensions.

Based on the classification results, item features in use before administration of the test have the highest prediction value in the outcome of the test, when variance from the two dimensions is considered important. Item features only have a high prediction rate in all configurations, because of the fact that they are computed from the performance data.

Dimensions	**Percentage correctly predicted**				
	1	**2**	**3**	**4**	**5**
All features	83,9	83,9	50	53,8	35,9
Word features	84,6	88,5	83,9	68,8	68,8
Item features (before adm)	100	100	83,9	68,8	68,8
Item features	94,9	100	97,3	94,9	97,4

TABLE 4.7. Classification accuracy in five cluster configurations.

The information of the usefulness of the independent variables is present collectively in the discriminant functions which are extracted in the analysis. They account for the discriminant structure in a similar fashion to principal component analysis: the first function describes most of the variance that can be accounted for by the use of the independent variables, the second function describes what is left of the variance after the first one has been extracted, and so on. The total information thus obtained from the variables is used to see if the grouping can be justified by the continuous variables. (Klecka 1990: 42). It is, however, important to test if the information in the discriminant functions is statistically substantial enough to make the same prediction justified with future tests.

Wilks' Lamda is a test used to assess the significance of the discriminant functions. It is important to look at the significance before any of the functions have been derived, because this tells about the usefulness of the features as a complete set. High significance suggests that the independent variables are important. (Klecka 1990: 41). The variable sets were assessed in the four situations, and the results can be seen in appendix 10.

All predictor variables were entered in the analysis together to see their usefulness in explaining the clusters of the response data. The significance of Wilks' lamda for the one dimensional solution was .090, two dimensional solution .000, three dimensional solution .002 and .020 for four and .412 for five dimensional solutions, which indicates that all but the last configurations make notable use of the predictor variables. The first discriminant functions, however, have the highest significance for two and three dimensional solutions .002 and .382, which shows that the first function can be used to generalise in the two dimensional case only. In this case Klecka (1973: 41) suggests emphasising substantial findings of the study, because the variable set as a whole has been noted to work well. This is a reasonable thing to do also in this study, because of the small number of items.

The results show that as the number of dimensions grow, word features loose their explanatory power of item clusters. The number of clusters also increases rapidly after the two dimensions corresponding to the amount of variance explained. Therefore it is reasonable to use either the one or two dimensional solution in explaining the classification results, in particular, as the stress values for the higher number of dimensions showed that the data may contain measurement error. Both of these configurations have a high prediction rate in terms of the prediction variables.

In the one dimensional configuration, the groups are divided best in terms of the number of distractors based on form, and inversely by the number of meaning based distractors and also by the weighted amount of the meaning based distractors. They are also distinguished by the number of distractors based on unique sources of psycholinguistic error sources. Most important word features relate to number of letters and the number of homonyms, and it seems that word length relates directly to the classification of the items.

In the two dimensional configuration, the most important features that can be used for classification are number of well behaving distractors (.525), which indicates that the separateness of the clusters relates to guessing as a prominent response strategy. Estimation error for the b-parameter correlates with this function (-.392) indicating that where estimation is most precise the discrimination of the items is highest. This is closely linked with the correlation with the b-parameter (.379), suggesting also that the difficulty of items is related to the distinctiveness of the clusters. The second distractor classification correlates also with this function. The Number of distractors based on the psycholinguistic errors is negatively related to the group differences (-.421), which indicates that the differences are greatest when items are based on few error sources. The number (.437) and the appeal (.305) of the distractors based on intralingual errors also correlate with this function. This shows that these distractors have an important effect, especially when the distractors have a high appeal.

5 CONCLUSION

The primary aim of this study was to find out which features of test items serve in the construction of a vocabulary size measurement scale. A limitation for the generalisability of the results was the fact that the items did not provide significant results. This may be due to a number of factors, and some of them were identified in this study. Most notably, a similar study would be required with a data set that allows positively defined tetrachoric correlations to be computed. The use of Pearson correlation matrix tends to underestimate the correlations between dichotomous items, and thus the true relationships between items are not shown as high as they may actually be.

An unbalanced sampling procedure was used for selecting the words to the tests. The process description used for constructing the items does not mention how to deal with situations in which the selection of target words becomes biased in some way. The recommendation based on this study is to specify with reference to the word qualities the features of words that are important for a fair measurement of vocabulary size, and include these into the definition of vocabulary. For instance, none of the classifications reviewed for this study take into account the representation of word classes in a language.

The structure of the vocabulary size test was examined with respect to the word and item features. The aims was to find out the internal structure of the test in terms of the features. A secondary aim was to investigate the relationship of the frequency counts in particular to find out the relationships between different types of word counts.

First, the analysis of the structure of the test showed that items were heavily based on nouns. This was due to a simple random selection of words for the items. If word classes, in fact, have an influence on the difficulty of the acquisition of words, it might be useful to develop a sampling system which takes into account the distribution of word classes in a language, and to draw samples that have a better representation of the word classes.

Second, the results of the analysis also indicated that the similarity of word forms, rank in the frequency count and number of homonyms were related to the difficulty of the items. It was also noted that difficulty is also related to the number of affixes the word contains. Item features that had the most significant effect were +the number of unattractive distractors in the items and guessing involved in responding to the item. It was also noted that the appeal of psycholinguistic error sources was not related to ~~the~~ item difficulty, while the classification based on the structure of the learner lexicon indicated that distractor types most likely influencing the difficulty of the items were form-based distractors and distractors for which it was difficult to perceive any linguistic grounds other than ~~a~~ general plausibility.

Test reliability was estimated in order to find out the extent to which subsequent scaling can rely on the instrument used. It was noted that the reliability was quite good, and that the items used in the test seem to form a rather consistent measurement instrument.

Item difficulty was computed by using two measures. The p-value was inversed, and it was used as a sample-dependent measure of difficulty. The b-parameter was used to compute sample-free measures of difficulty. The difficulty scale based on the b-parameter can be seen in appendix 7. It shows that the items seem to cluster at a certain point in the scale, and that the measurement is concentrated at the centre part of the items, with a few extreme, easy items towards the top of the scale.

Item difficulty is probably the most important aspect attributed to test items. The importance of difficulty became evident in the fact that it correlates with some of the word features, but also in the fact that the results of multi-dimensional scaling suggests that one of the major dimensions in the test is related to item difficulty.

The estimation of the usefulness of items in terms of performance was examined with p-values in the sex and age groups. The differences between sexes sums up that the test that is more favourable to women than men. This may probably also be due to a more sizable sample of women taking part in the test. On the whole, the differences were observed for words which were considered important and concrete. The result seems to be slightly problematic, because on the basis of previous research, these words are also considered important on this level of proficiency.

The results for the age groups show that the differences between the group of the youngest test subjects and the other two groups is pronounced. The proportion correct in the two groups with older subjects shows a lower total difference between those test subjects than in the group of the youngest subjects. Group differences seem to be greatest between the oldest examinees and the youngest exminees. The differences were seen in items which had a number of useful distractors. This may not be due to a bias towards ~~to~~ the older persons, but rather it shows that the effect in performance can be related to the age of the examinees.

Item discrimination was investigated to see if the items distinguish well between persons of low and high ability. Two indexes were used which indicate that the items discriminate reasonably well. The two indexes were more closely associated with the b-parameter estimate of item difficulty. The discrimination of the vocabulary items was low. The rule of thumb for interpreting item discrimination is that for the items to be good, the discrimination index should be above .300. Based on the point-biserial correlation 18 items would have passed. However, only two items would have been considered poor, because their biserial correlation was below .100, which may be a more practical limit for item discrimination, especially since vocabulary tests measure a concept that is highly variable between examinees.

Three methods were used to assess the dimensionality of the performance data. A rough estimate of the inter-item correlations suggested that the test is comprised of one dimension. Investigation of the eigenvalues of principal components of the data also indicated that the performance data are by and large unidimensional. Four other components were found with considerably less importance compared to the first one. Finally, the stress values of the multi-dimensional scaling were investigated. The stress values indicate that one dimension is most likely, although it was also found that this dimension contains a considerable degree of measurement error.

The value of these findings was decreased by the fact that a proper correlation coefficient could not be used for the principal component analysis. Pearson correlation, which was used, showed that a unidimensional solution could be supported. Investigation of the clusters, however, indicated that it may be more useful to try to interpret the data through two dimensions, even though the second dimension seems to be related to measurement accuracy.

The clustering of the items was investigated by hierarchical clustering of the coordinates of the multi-dimensional space in two cases. Items clearly formed clusters up to a three dimensional solution, and after that the items were included in several groups. Two large clusters were observed in both the one dimensional and two dimensional cases, and several smaller clusters. In the one dimensional space clusters were more clearly apart from each other, while in the two-dimensional space some of the clusters were located closer to the large cluster.

- ed

It was the assumption that the groups based on performance data would be related to response strategies employed by the test subjects. The usefulness of word and item features to explain clustering was investigated with discriminant analysis. The results show that the word features were quite good predictors of the clusters, but at the same time the results were not statistically significant. Most prominent features explaining the differences between items were the number of homonyms and, in the one dimensional configuration, the number of letters. The correlations, however, were modest. Item features were also found to predict results extremely well, and the results were statistically significant. In the one dimensional solution the items were more clearly apart from each other based on the first classification of distractors emphasising the number of meaning and form-based distractors.

It was also evident from the results that the classification had a high rate of predicting the results correctly. When the structure of the data was more clearly related to the measurement of skills and was free of error, prediction was successful, but as the amount of error increases, it is the item features used as a single criterion which predict correctly the results.

Despite the problems, it was found that reliable and general predictions of item performance can be made on the basis of item features. The test had a good reliability, and items tended to discriminate well according to the ability of the examinees. The test structure was investigated, and the test was considered effective for the assessing vocabulary size. Item descriptors were assessed in order to find which features are important for the item classification according to cognitive demand. The features in general formed a useful set of descriptors. The fact that these descriptors worked well suggests that they would be more useful in the specification of test construction. It is useful, then, to incorporate ~~the~~ word features in all the phases of item construction, including word sampling, in order to construct a balanced test. It would also be useful to use the features examined in this study in identifying good and bad specimen items for test specifications. The fact that none of the word features were important in explaining either the test structure means that they cannot be used in scale construction as separate variables. Features that may be of interest in future studies, however, include the rank of word in frequency lists and number of homonyms.

Also, two different perspectives on estimating the kinds of errors that are involved in distractor preparation were considered. A classification based on the structure of the learner lexicon seemed to be more useful in explaining what kind of distractors are useful for item construction on the intermediate level. The classification based on psycholinguistic error sources seemed to indicate that each of the error sources is ~~as~~ just as valid for assessment purposes on this proficiency level. It was, however, significant in explaining the cognitive demand that the items pose on the examinee. This classification may have more importance in the research of the cognitive strategies employed by the subjects in responding to the tasks. However, a combined use of these two classifications seems to be useful in assessing the quality of tests and items.

In light of the results of this study, it would be also useful to study the usefulness of the word and item features also on other levels of proficiency. Especially as the results suggest that the composition of learner lexicon has importance in the quality of the items, it would be useful to examine the usefulness of the error classifications on the other proficiency levels and across the different languages of the FNC assessment system.

BIBLIOGRAPHY

Anderson, Richard C. and Peter Freebody (1983). Reading comprehension and the assessment and acquisition of word knowledge. In *Advances in reading / language research*, 2, 231-256.

Anderson, Richard C. and Willian E. Nagy (1984). How many words are there in printed school English? *Reading research quarterly* 19, 304-330.

Arabski, Janusz (1993). A Foreign Lexis Acquisition Model. Ed. Wolfgang Kühlwein. In *Language as a structure and language as a process.* Wissenschaftlicher Verlag Trier.

Avokari, Marja-Liisa, Terttu Hirvenoja and Ronnie Wallace (1980). *Words onwards: englannin keskeistä sanastoa.* Juva: WSOY.

Avokari, Marja-Liisa, Terttu Hirvenoja and Ronnie Wallace (1997). *Word files.* Porvoo: WSOY.

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

Bachman, Lyle and Adrian Palmer (1996). *Language testing in practice.* Oxford: Oxford University Press.

Biber, Douglas (1993). Representativeness in corpus design. *Literary and linguistic computing* 8(4), 243-257.

Bourne, Lyle F., Roger Dominowski, Elizabeth F. Loftus and Alice F. Healy (1986). *Cognitive processes.* Second edition. Englewood Cliffs, N J: Prentice Hall.

Buck, Gary (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language testing* 145-170.

Carroll, John B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika* 26, 347-372.

Carter, Ronald (1987). *Vocabulary: applied Linguistic Perspective.* London: Unwin Hyman.

Carter, Ronald and Michael McCarthy (1988). *Vocabulary and Language Teaching.* London: Longman.

Choi, Inn-Chull and Lyle F. Bachman (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language testing* 9, 51-78.

Cohen, Andrew D. (1990). *Language learning: insight for learners, teachers and researchers*. Boston, MA: Heinle & Heinle.

Collins Cobuild (1996). *Lemmatised frequency word list.* Available FTP: ftp.itri.bton.ac.uk Directory: pub/bnc/lemma.doc.

Crystal, David (1987). *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

De Villiers, Jill G and Peter A. de Villiers (1978). *Language acquisition.* Cambridge: Harvard University Press.

Dunn-Rankin, P. (1988). Scaling Methods. In *Educational research, methodology, and measurement: an international handbook.* Ed. John P. Keeves. Oxford: Pergamon Press.

Ellegård, Alvar (1960). Estimating vocabulary size. *Word*, 16, 219-244.

Ellis, Nick C. and Alan Beaton (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language learning*, 43, 559-617.

Ellis, Rod (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Hambleton, Ronald K. (1985). *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, Ronald K. (1988). Criterion-referenced measurement. In *Educational research, methodology, and measurement: an international handbook.* Ed. John P. Keeves. Oxford: Pergamon Press

Hambleton, Ronald K. (1993). Principles and selected applications of item response theory. In *Educational research.* Ed. Robert L. Linn. Phoenix: Oryx Press.

Hammerly Hector (1982). *Synthesis in SL teaching: an introduction to linguistics.* Preliminary edition. Blaine, WA: Second Language Publications.

Harris, Richard J. Harris (1975). *A primer of multivariate statistics.* New York: Academic Press.

Henning, Grant (1973). Remembering foreign language vocabulary: acoustic and semantic parameters. *Language learning*, 23, 185-196.

Henning, Grant (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. *Language testing* 5, 83-99.

Herman, J. L. (1990). Item writing techniques. In *The international encyclopedia of educational evaluation*. Ed. Herbert J. Walberg and Geneva D. Haertel. Oxford: Pergamon Press.

Higa, Masanori (1965). The psycholinguistic concept of "difficulty" and the teaching of foreign language vocabulary. *Language learning*, 15, 167-179.

Jackson, Howard (1991). *Words and their meaning*. London: Longman.

Johansson and Hofland (1989). *Frequency analysis of English vocabulary and grammar, volume 1*. Oxford: Clarendon Press.

Kendall, Maurice G. (1967). *The advanced theory of statistics II: inference and relationship*. Second edition. London: Charles Griffin.

Kilgarriff, Adam (1996). *Frequency list: accompanying documentation*. Available FTP: ftp.itri.bton.ac.uk Directory: pub/bnc/README.

Kim, Jae-On and Charles W. Mueller (1978). *Factor analysis: statistical methods and practical issues*. London: Sage.

Klecka, William R. (1990). *Discriminant analysis*. London: Sage.

Konttinen, Raimo (1981). *Testiteoria: johdatus kasvatus ja käyttäytymistieteellisen mittauksen teoriaan*. Helsinki: Oy Gaudeamus Ab.

Kruskal, Joseph B. and Myron Wish (1978). *Multidimensional scaling*. London: Sage.

Lado, R. (1957). *Linguistics across cultures*. Ann Arbor: University of Michigan Press.

Laufer, Batia (1989). A factor of difficulty in vocabulary learning: deceptive transparency. In *Vocabulary acquisition*, ed. Paul Nation and Ron Carter. Amsterdam: Free University press.

Luria, Alexander R. (1982). *Language and cognition*. Washington, DC: V.H, Winston & Sons.

Lyons, John (1968). *Introduction to theoretical linguistics.* Cambridge: Cambridge University Press.

Lyons, John (1987). *Semantics 1-2.* Cambridge: Cambridge University Press.

Marascuilo, Leonard A. and Joel R. Levin (1983). *Multivariate statistics in the social sciences: a researchers guide.* Monterey, California: Brooks/Cole.

Meara, Paul (1996). The dimensions of lexical competence. In *Performance and competence in second language acquisition.* Ed. G. Brown, K. Malmkjaer and J. Williams Cambridge: Cambridge University Press.

Miettinen, Eino and Pirkko Uotinen (1983). *Nyky-englannin keskeinen sanasto.* Helsinki: Otava.

Na, Liu and Nation, I.S.P (1985). Factors affecting guessing vocabulary in context. *RELC Journal* 16(1), 33-42.

Nation, I.S.P (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, Paul (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language*, 8(2), 689-696.

Nation, Paul (1993). Using dictionaries to estimate vocabulary size: essential, but rarely followed, procedures. *Language testing*, 10, 27-40.

Nummenmaa, Tapio, Raimo Konttinen, Jorma Kuusinen and Esko Leskinen (1996). *Tutkimusaineiston analyysi.* Helsinki: WSOY.

Quirk, Randolph and Sidney Greenbaum (1973). *A university grammar of English.* Essex: Longman.

Palmberg, Rolf (1988). On lexical inferencing and language distance. *Journal of pragmatics* 12, 207-214.

Perkins, Kyle (1987). A construct definition study of a standardised vocabulary test. *Language testing*, 4, 125-141.

Piedmont, Ralph L. and Michael E. Hyland (1993). Inter-item correlation frequency distribution analysis: a method for evaluating scale dimensionality. *Educational and psychological measurement* 53, 367-378.

Pressley, Michael and Elizabeth G. Ghatala (1988). Delusions about performance on multiple-choice comprehension tests. *Reading research quarterly* 23, 454-464.

Rasch, George (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Richards, I. A. (1943). *Basic English and its uses*. London: Kegan Paul.

Rosenbaum, Paul R. (1988). Item bundles. *Psychometrika* 53 (3), 349-359.

Röman, Kyllikki, Juhani Karvonen, Annika Takala, Oiva Ylinentalo (1970). *Opettajan sanastokirja*. Jyväskylä: Gummerus.

Schwanenflugel, Paula and Randall Stowe (1989). Context availability and the processing of abstract and concrete words in sentences. *Reading research quarterly* 24(1), 114-127.

Schwanenflugel, Paula and Carolyn Akin (1994). Developmental trends in lexical decisions for abstract and concrete words. *Reading research quarterly* 29(3), 251-264.

Simonsen, Stephen and Harry Singer (1992). Improving reading instruction in the content areas. In *What research has to say about reading instruction*. Ed. S Jay Samuels and Alan E. Farstrup. Newark, DE: International Reading Association.

Spolsky, Bernard (1989). *Conditions for second language learning*. Oxford: Oxford University Press.

Tabachnick, Barbara and Linda S. Fidell (1989). *Using multivariate statistics*. Second Edition. New York: Harper & Row.

Takala, Sauli (1984). *Evaluation of students' knowledge of English vocabulary in the Finnish comprehensive school*. Jyväskylä: University of Jyväskylä.

Thorndike, R. L. (1973). Reading as reasoning. *Reading research quarterly* 9:135-147.

Underwood, B. J. (1969). Attributes of memory. *Psychological review* 76, 559-573.

U.S. Department of Health, Education, and Welfare (1974). *The rationale, development, and standardization of a basic word vocabulary test*. Rockville, MD: National Center for Health Statistics.

Verhelst, Norman D., C.A. Glas and H. H. F. M.Verstralen (1995). *One-parameter logistic model: OPLM*. CITO: Arnhem.

Verhelst, Norman D. (1998). *The utility OPCOR*. Tentative operating manual. CITO: Arnhem.

White, Lydia (1989). *Universal grammar and second language acquisition.* Amsterdam: John Benjamins.

Wood, R. (1988). Item analysis. In *Educational research, methodology, and measurement: an international handbook.* Ed. John P. Keeves. Oxford: Pergamon Press.

Wood, Robert (1991). *Assessment and Testing.* Cambridge: Cambridge University Press.

Woods, Anthony (1983). Principal components and factor analysis in the investigation of the structure of language proficiency. In *Current developments in language testing.* Ed: Arthur Hughes and Don Porter. London: Academic Press.

Wysocki, Katherine E. And Joseph R. Jenkins (1987). Deriving word meanings through morphological generalisation. *Reading research quarterly* 22(1), 66-81.

Appendix 1:

The multiple choice test used in this study

Instructions:

Seuraavassa on 40 englannin sanaa, jotka on valittu sattumanvaraisesti muutamien tuhansien sanojen joukosta. Siksi joukossa todennäköisesti on myös sanoja, joita et tunne. Älä anna tämän häiritä itseäsi. Valitse kussakin kohdassa mielestäsi paras suomenkielinen vastine ja mustaa (●) vastineen kirjain (a, b, c tai d) optisen lomakkeen kohtiin M1 - M40.

Test:

M1. ache	a. alue	b. kipu	c. väite	d. tuhka
M2. association	a. kilpailu	b. yhdistys	c. vieraantuminen	d. syytös
M3. commercial	a. kaupallinen	b. koominen	c. tuleva	d. vasemmisto
M4. desk	a. laatikko	b. pöytä	c. taulu	d. lipas
M5. dandruff	a. voikukka	b. vuoristo	c. keimailija	d. hilse
M6. estate	a. tila	b. muuri	c. väittää	d. estää
M7. edge	a. aita	b. viha	c. kutina	d. reuna
M8. element	a. osa	b. puoli	c. kaasu	d. alku
M9. foam	a. kupla	b. kumi	c. muovi	d. vaahto
M10. form	a. vaahdota	b. temppuilla	c. muodostua	d. seota
M11. grease	a. hölmö	b. rasva	c. ongelma	d. possu
M12. gutless	a. avuton	b. uljas	c. pelkuri	d. lattea
M13. hunting	a. hunnutus	b. pikamatka	c. metsästys	d. aivastus
M14. income	a. tulovero	b. tuontitulli	c. tulot	d. aula
M15. immoderate	a. modeemiton	b. iätön	c. kohtuuton	d. muodoton
M16. jelly	a. masu	b. hyytelö	c. hillo	d. kudos
M17. messenger	a. metalliseos	b. sekottaja	c. lähetti	d. hieroja
M18. molar	a. kuuhullu	b. poskihammas	c. valas	d. home
M19. mean	a. ilkeä	b. keskiaika	c. paisti	d. pääosa
M20. numerous	a. lukematon	b. kokonainen	c. numeroitu	d. lukuisa
M21. oil-well	a. rasvaus	b. hiushoito	c. lahjonta	d. öljylähde
M22. plot	a. juoni	b. aitta	c. näyte	d. solmu
M23. pond	a. sidos	b. osake	c. ponnahdus	d. lampi
M24. presume	a. savustaa	b. olettaa	c. mainostaa	d. varmistaa
M25. reality	a. kiinteistö	b. todellisuus	c. kokojyvä	d. kelaus

M26. roller	a. rata	b. tela	c. maanvyörymä	d. potku
M27. rookie	a. kukko	b. kiertotie	c. lukko	d. alokas
M28. riddle	a. ristisana	b. arvoitus	c. palapeli	d. arvaus
M29. rust	a. talonpoika	b. sorto	c. kuori	d. ruoste
M30. staff	a. aine	b. täyte	c. henkilökunta	d. teippi
M31. seam	a. savupiippu	b. syvänne	c. sauma	d. suoja
M32. supply	a. varasto	b. tarjous	c. avustus	d. uhraus
M33. Sunday	a. lauantai	b. sunnuntai	c. maanantai	d. tiistai
M34. turn grey	a. kalveta	b. homehtua	c. harmaantua	d. tulla vihaiseksi
M35. there	a. tuolla	b. tässä	c. tänne	d. täällä
M36. treasure	a. petos	b. lipas	c. kirstu	d. aarre
M37. untidy	a. sitomaton	b. kiireellinen	c. sotkuinen	d. vetoinen
M38. villager	a. roisto	b. päällystakki	c. kyläläinen	d. keritsijä
M39. a while ago	a. ennen taukoa	b. vastikään	c. katkelma	d. pari vuotta sitten
M40. ward	a. osasto	b. sarana	c. varoitus	d. holhooja

Appendix 2:

Word class and semantic classification of words according to three sources of learner lexicon

ache N

WOW:	social services/symptoms and illness
WF:	health and illness/symptoms
EKS:	health, illness, life and death/illness

association N

WOW:	economic life/economic groupings
WF:	where to stay/overnight accommodation economy/economic life economy/international co-operation Christianity/church organisations
EKS:	N/A

commercial A

WOW:	work/business and industry (adj) advertising/sales promotion (n) communications/mass media sport/ball games
WF:	let's go outdoors/fishing media and culture/television and radio economy/economic life
EKS:	business, industry, etc./general terms

dandruff N

WOW:	N/A
WF:	N/A
EKS:	N/A

desk N

WOW:	housing/furniture and household equipment
WF:	education and research/school furniture and equipment
EKS:	education/reading and writing

edge N

WOW:	science/measures
WF:	N/A
EKS:	quantities, measure and forms: measures

element N

WOW:	science/chemistry
WF:	fine arts/good design
EKS:	nature and the Universe/general terms

estate N

WOW:	work/different occupations
WF:	personal details/death housing/place to live (/homes and house hunting) employment/occupations economy/producers traffic/road traffic
EKS:	plants and farming/farming

foam N

WOW:	N/A
WF:	N/A
EKS:	N/A

form V		
	WOW:	work/applying for a job
		the arts/painting and sculpture
		education/school attendance (BrE)
	WF:	education/higher education
		earning one's living/job-hunting
		meetings and conferences/forms of meeting
		travelling/international travel
		fine arts/paintings
		fine arts/good design
	EKS:	quantities, measures and forms/the form of things
grease N		
	WOW:	N/A
	WF:	dressing/beauty products
	EKS:	metals, minerals and other materials
gutless N		
	WOW:	N/A
	WF:	N/A
	EKS:	N/A
hunting N		
	WOW:	N/A
	WF:	let's go outdoors/hunting
	EKS:	(hunt) animals and animal life: wild animals
income N		
	WOW:	work/business and industry
		work/condition, pay and taxes
	WF:	earning one's living/pay and taxation
	EKS:	business, industry, etc/wages and earning
immoderate A		
	WOW:	N/A
	WF:	N/A
	EKS:	N/A
jelly N		
	WOW:	N/A
	WF:	N/A
	EKS:	food, meals, etc.: food and cooking
mean A		
	WOW:	the human mind/character and appearance (adj.)
	WF:	human mind/character (adj)
		personal details/will and intentions (v)
	EKS:	human mind: will (v)
messenger N		
	WOW:	N/A
	WF:	N/A
	EKS:	language/correspondence and communication
molar N		
	WOW:	N/A
	WF:	personal details/human body
	EKS:	N/A
numerous A		
	WOW:	N/A
	WF:	N/A
	EKS:	education/mathematics
oil-well N		
	WOW:	nature/natural resources
	WF:	N/A
	EKS:	N/A

plot N		
	WOW:	the arts/general vocabulary
	WF:	literature/fiction or non-fiction
	EKS:	building and home/building
pond N		
	WOW:	georgraphy/general terms
	WF:	let's go outdoors/call of the wild
		environment/physical geography
	EKS:	geography and landscape/general terms
presume V		
	WOW:	N/A
	WF:	N/A
	EKS:	N/A
reality N		
	WOW:	N/A
	WF:	N/A
	EKS:	other words
riddle N		
	WOW:	N/A
	WF:	N/A
	EKS:	the human mind/intelligence
roller N		
	WOW:	N/A
	WF:	N/A
	EKS:	sport
rookie N		
	WOW:	N/A
	WF:	N/A
	EKS:	N/A
rust N		
	WOW:	N/A
	WF:	N/A
	EKS:	metals, minerals and other materials
seam N		
	WOW:	N/A
	WF:	N/A
	EKS:	N/A
staff N		
	WOW:	education/school attendance
		holidays and travel/travelling
	WF:	where to stay/facilities and services
	EKS:	education/education and school
Sunday N		
	WOW:	religions and beliefs
		daily routine/sleeping and waking
	WF:	general vocabulary/time
		christianity/christian life
	EKS:	time/year and its divisions
supply N		
	WOW:	economic life/business and economy
		public services/other services
		technology/space research
	WF:	economy/economic life
	EKS:	business, industry, etc/buying and selling
there P		
	WOW:	(Appears in exemple sentences in several categories.)
	WF:	N/A
	EKS:	other words

treasure N

WOW:	N/A
WF:	N/A
EKS:	business, industry, etc./property and money

turn grey VP

WOW:	N/A
WF:	N/A
EKS:	(turn) other words

untidy A

WOW:	daily routine/household chores
WF:	dressing/at the hairdresser's
EKS:	other words

villager N

WOW:	the structure of society/population
WF:	N/A
EKS:	(village) geography and landscape: general terms

a while ago Adv

WOW:	N/A
WF:	(while) general vocabulary/conjunctions
EKS:	(while) time/general terms

ward N

WOW:	social services/health services
WF:	health and illness/at the hospital
EKS:	health, illness, life and death: illness

Abbreviations used for word classes:

A	= Adjective
Adv	= Adverbial
N	= Noun
P	= Pronoun
V	= Verb
VP	= Verb Phrase

Abbreviations used for sources of semantic classification

WOW	= Words Onwards
WF	= Word files
EKS	= Englannin keskeinen sanasto

Appendix 3:

Content categorisation of target words in the multiple choice items.

gutless	character (1.1)	human (1)
mean		
molar	body (1.2)	

oil-well	resources (2.1)	nature (2)
pond	landscape (2.2)	

ache	illness (3.1.1)	personal (3.1)
dandruff	hygiene (3.1.2)	
income	work (3.2.1)	organisation (3.2)
staff		
hunting	leisure (3.2.2)	
riddle	education (3.2.3)	
rookie	status (3.2.4)	
villager		
association	general (3.2.5)	
untidy		
commercial	business (3.2.6)	
supply		
treasure		
messenger	services (3.2.7)	
desk	furniture (3.3.1)	equipment (3.3)
ward	housing (3.3.2)	
estate		
seam	clothes (3.3.3)	
roller	general (3.3.4)	
plot	general (3.4.1)	arts (3.4)

edge	form (4.1)
form	
element	object (4.2)
reality	
foam	material (4.2.1)
grease	
jelly	
rust	
numerous	quantities (4.3)
Sunday	time (4.3.1)
a while ago	
immoderate	
presume	
there	
turn grey	

Appendix 4:

Variables used in this study.

Variable name	Word / item feature	Scale type
abs_con	Abstractness of the word	Binary
base_dis	Number of affixes / transformations from base word	Count
basic	Basic word	Binary
content	Semantic fields defined in LOB frequency dictionary	Nominal
context	Number of context areas where word was encountered in LOB corpus	Count
distnc	Similarity of English and Finnish word forms	Scale
freq_BNC	Frequency in BNC corpus	Count
freq_Fin	Frequency of the Finnish word (correct choice)	Count
freq_LOB	Frequency in LOB corpus	Count
freq_rank	Rank in BNC / CC frequency word list	Count
imgrated	Imageability (.7469)	Scale
imprtnc	Importance for the language learner (.8702)	Scale
level	Level on which the word is usually taught	Scale
no_hom	Number of homonyms	Count
no_let	Number of letters	Count
semfld1	Precise semantic field	Nominal
semfld2	Broad semantic field	Nominal
wordclas	Word class	Nominal
#CLAS1_1	Number of distractors for error classification 1: form	Count
#CLAS1_2	Number of distractors for error classification 1: meaning	Count
#CLAS1_3	Number of distractors for error classification 1: plausibility	Count
#CLAS2_1	Number of distractors for error classification 2: transfer	Count
#CLAS2_2	Number of distractors for error classification 2: intralingual	Count
#CLAS2_3	Number of distractors for error classification 2: unique	Count
no_CLAS1	Number of distractors used in item based on error classification 1	Count
no_CLAS2	Number of distractors used in item based on error classification 2	Count
age1_2	Difference in performance between age groups 1 and 2	Scale
age1_3	Difference in performance between age groups 1 and 3	Scale
age2_3	Difference in performance between age groups 2 and 3	Scale
alpha	Difference between alpha and alpha if item removed from test	Scale
gendiff	Difference in performance between the sexes	Scale
inv_p	Inversed classical difficulty value (inverse p-value)	Scale
itTotSPS	Biserial correlation (SPSS)	Scale
itemattr	Number of distractors selected by less than 5% of the subjects	Count
itTotOPL	Point-biserial correlation (OPLM)	Scale
OPLMDiff	Item difficulty parameter value (OPLM)	Scale
OPLMSE	Standard error of the OPLMDiff estimate	Scale
#WCLAS1_1	Attraction of distractors based on #CLAS1_1 and inv_p	Scale
#WCLAS1_2	Attraction of distractors based on #CLAS1_2 and inv_p	Scale
#WCLAS1_3	Attraction of distractors based on #CLAS1_3 and inv_p	Scale
#WCLAS2_1	Attraction of distractors based on #CLAS2_1 and inv_p	Scale
#WCLAS2_2	Attraction of distractors based on #CLAS2_2 and inv_p	Scale
#WCLAS2_3	Attraction of distractors based on #CLAS2_3 and inv_p	Scale
tiagrafc	Correct choice that works well	Binary
tiagrad	Number of distractors that work well	Count

Appendix 5:

Statistics from the reliability analysis.

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
OM01	30.4526	24.5436	.4221	.7975
OM02	30.2758	25.8331	.2662	.8038
OM03	30.2358	26.2734	.1820	.8061
OM04	30.3621	25.8897	.1332	.8071
OM05	30.6905	25.0369	.2378	.8050
OM06	30.5663	24.4782	.3772	.7989
OM07	30.3074	25.4749	.3331	.8018
OM08	30.2463	26.2409	.1446	.8062
OM09	30.2484	26.0057	.2905	.8043
OM10	30.2484	26.2377	.1388	.8063
OM11	30.4063	24.7523	.4103	.7983
OM12	30.8063	24.4561	.3632	.7995
OM13	30.2358	26.3114	.1456	.8064
OM14	30.2695	25.8175	.2939	.8035
OM15	30.5642	24.6810	.3332	.8008
OM16	30.5095	25.2125	.2334	.8047
OM17	30.3958	25.1975	.3004	.8020
OM18	30.7937	25.3329	.1795	.8075
OM19	30.4779	24.2121	.4843	.7949
OM20	30.5684	25.5074	.1544	.8081
OM21	30.4274	24.8275	.3708	.7995
OM22	30.7347	24.5497	.3376	.8007
OM23	30.6358	24.3924	.3782	.7989
OM24	30.3895	24.8839	.3927	.7990
OM25	30.2337	26.2554	.2248	.8059
OM26	30.3958	25.4548	.2311	.8043
OM27	30.4695	25.5703	.1654	.8070
OM28	30.5937	24.1109	.4492	.7958
OM29	30.4063	25.0645	.3268	.8011
OM30	30.3558	25.3352	.3021	.8021
OM31	30.5937	24.8704	.2846	.8029
OM32	30.7011	24.2944	.3916	.7983
OM33	30.2358	26.2818	.1739	.8062
OM34	30.3853	26.1403	.0532	.8099
OM35	30.2800	26.0839	.1456	.8061
OM36	30.3579	24.7619	.4727	.7970
OM37	30.5326	24.3465	.4200	.7972
OM38	30.2947	26.1999	.0805	.8076
OM39	30.2926	25.8825	.2072	.8049
OM40	30.8084	24.8767	.2750	.8033

Appendix 6:

Differences in item performance in age and sex groups.

Item	Groups 1-2	Groups 2-3	Groups 1-3	Women-men
1	-.005	-.011	-.016	.070
2	-.028	-.017	-.045	-.070
3	-.020	-.009	-.029	-.020
4	-.079	.009	-.070	.020
5	-.128	-.011	-.139	-.020
6	.004	.169	.173	-.150
7	-.117	-.006	-.123	-.080
8	-.040	.011	-.029	-.020
9	.017	-.001	.016	.000
10	-.002	.015	.013	.010
11	-.248	.104	-.144	-.130
12	-.073	-.149	-.222	-.020
13	-.005	.015	.010	.000
14	-.051	.019	-.032	-.020
15	-.018	.091	.073	-.030
16	-.072	.053	-.019	.090
17	.003	-.057	-.054	-.060
18	-.082	.047	-.035	-.030
19	-.208	-.111	-.319	.010
20	.069	.106	.175	-.040
21	.017	.009	.026	-.100
22	.091	-.009	.082	.070
23	.113	.178	.291	-.020
24	-.114	.028	-.086	-.040
25	-.024	-.005	-.029	-.030
26	.039	.009	.048	-.050
27	-.067	-.114	-.181	-.110
28	-.196	-.022	-.218	-.080
29	-.058	.005	-.053	-.100
30	-.071	.093	.022	.000
31	.054	-.023	.031	-.090
32	.024	-.045	-.021	-.110
33	-.034	-.005	-.039	.000
34	.053	-.028	.025	.060
35	-.015	-.037	-.052	-.020
36	-.065	-.038	-.103	-.070
37	-.043	-.068	-.111	.020
38	-.010	-.002	-.012	.030
39	-.053	.034	-.019	.010
40	-.033	-.242	-.275	.012

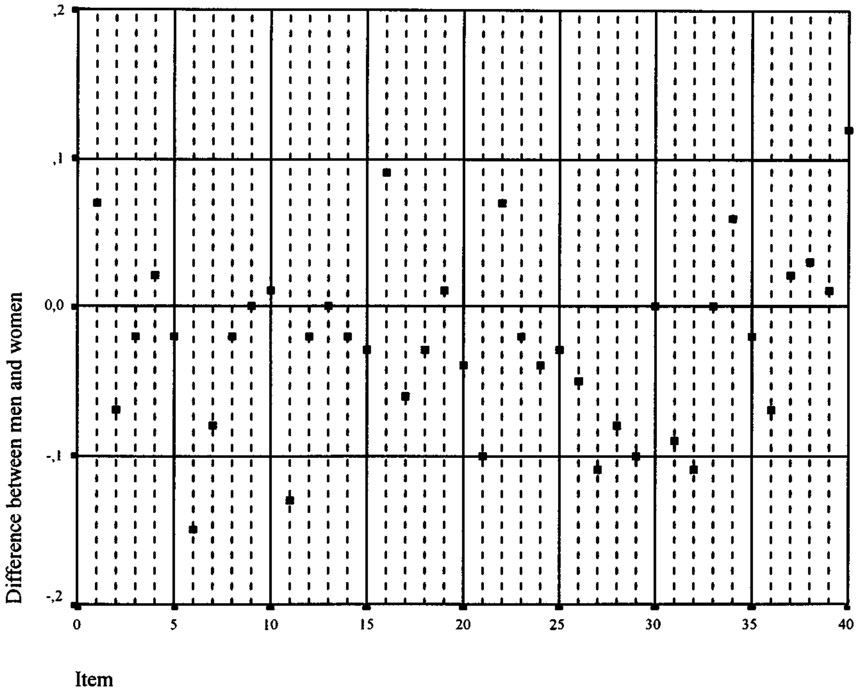


FIGURE A6.1. Differences in mean performance to items between men and women.

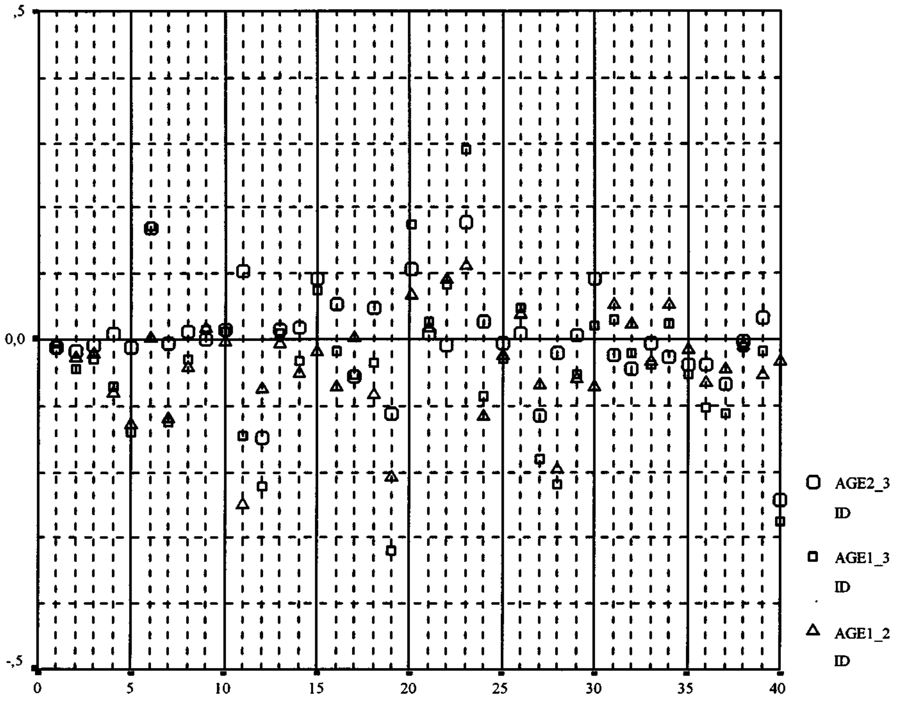
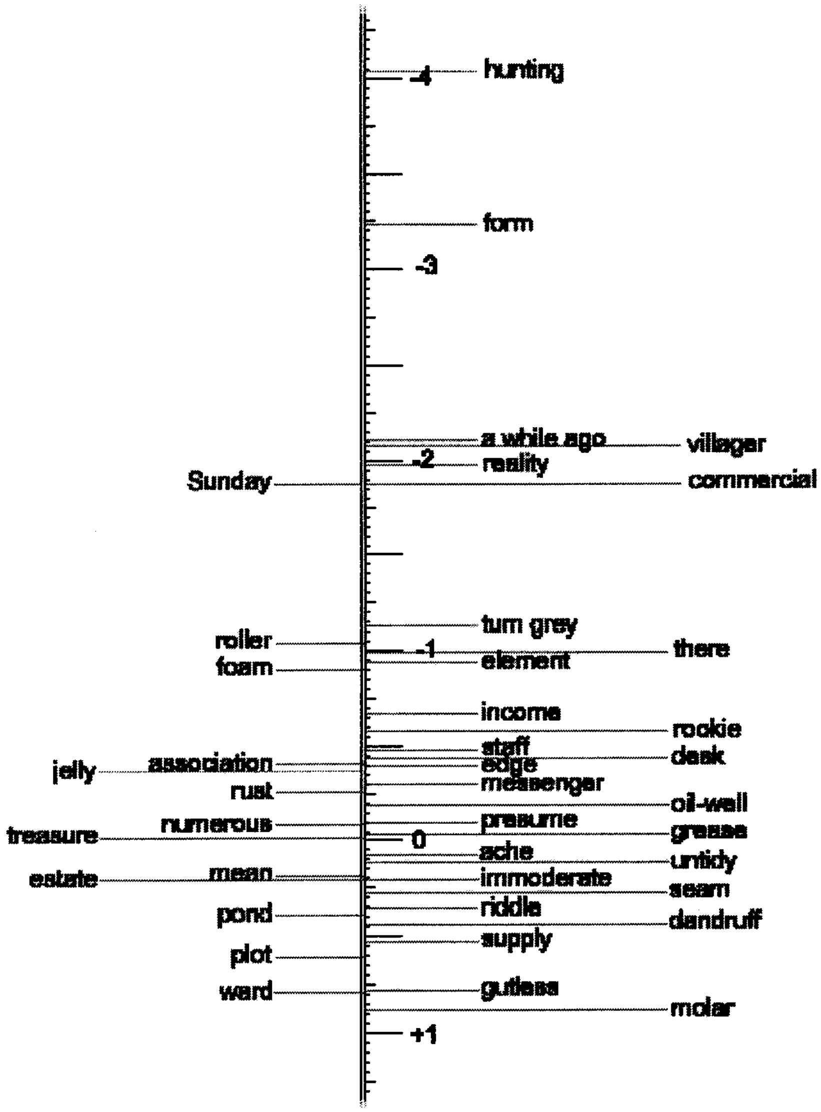


FIGURE A6.2. Differences in mean performance in the three age groups.

Appendix 7:

Item difficulty scale based on b-parameters.



Appendix 8:

Distribution of biserial and point-biserial correlations.

Item	R_{bis}	$R_{p\text{-}bis}$
1	.4221	.4880
2	.2662	.3060
3	.1820	.2010
4	.1332	.1990
5	.2378	.3280
6	.3772	.4550
7	.3331	.3800
8	.1446	.1720
9	.2905	.3170
10	.1388	.1670
11	.4103	.4720
12	.3632	.4450
13	.1456	.1650
14	.2939	.3300
15	.3332	.4140
16	.2334	.3150
17	.3004	.3660
18	.1795	.2720
19	.4843	.5480
20	.1544	.2440
21	.3708	.4370
22	.3376	.4220
23	.3782	.4590
24	.3927	.4530
25	.2248	.2420
26	.2311	.3000
27	.1654	.2460
28	.4492	.5230
29	.3268	.3930
30	.3021	.3610
31	.2846	.3700
32	.3916	.4720
33	.1739	.1930
34	.0532	.1240
35	.1456	.1890
36	.4727	.5230
37	.4200	.4930
38	.0805	.1300
39	.2072	.2540
40	.2750	.3630

Appendix 9:
Correlations

Word features

	Abs_con	base_dis	basic	context	distnc	freq_bnc	freq_fin	freq_lob	frq_cc	imgrated	imprtnc	level	no_hom	no_let
Abs_con	1.000	.398	-.008	.255	.025	.209	-.129	.227	.030	.429	-.514	-.154	.280	.131
Base_dis	.398	1.000	-.136	-.083	-.017	-.167	-.124	-.168	.131	.089	-.117	-.012	-.030	.273
Basic	-.008	-.136	1.000	.197	-.039	.168	.260	.160	-.313	-.268	-.309	-.302	-.033	-.616
Context	.255	-.083	.197	1.000	-.407	.386	.347	.366	-.671	.109	-.701	-.019	-.189	.081
Distnc	.025	-.017	-.039	-.407	1.000	.064	-.391	.085	.400	-.095	.276	-.043	.375	-.138
Freq_bnc	.209	-.167	.168	.386	.064	1.000	.001	.995	-.226	.317	-.391	-.351	-.217	-.069
Freq_fin	-.129	-.124	.260	.347	-.391	.001	1.000	-.010	-.330	.096	-.132	.358	-.259	.016
Freq_lob	.227	-.168	.160	.366	.085	.995	-.010	1.000	-.224	.335	-.370	-.362	-.197	-.079
Frq_cc	.030	.131	-.313	-.671	.400	-.226	-.330	-.224	1.000	-.058	.566	-.080	.190	.035
Imgrated	.429	.089	-.268	.109	-.095	.317	.096	.335	-.058	1.000	-.315	.239	.212	.383
Imprtnc	-.514	-.117	-.309	-.701	.276	-.391	-.132	-.370	.566	-.315	1.000	.099	.029	-.010
Level	-.154	-.012	-.302	-.019	-.043	-.351	.358	-.362	-.080	.239	.099	1.000	-.087	.372
No_hom	.280	-.030	-.033	-.189	.375	-.217	-.259	-.197	.190	.212	.029	-.087	1.000	-.106
No_let	.131	.273	-.616	.081	-.138	-.069	.016	-.079	.035	.383	-.010	.372	-.106	1.000
It_frm	-.022	-.014	-.211	-.175	.213	-.211	-.224	-.231	-.189	.179	-.042	.333	.175	.228
It_mng	-.083	-.173	.176	.276	-.282	.386	.171	.359	-.214	-.243	-.182	-.293	-.305	-.060
It_nul	.130	.236	.003	-.157	.132	-.246	.024	-.202	.465	.118	.278	-.007	.200	-.169
It_cl2_1	.077	.080	.243	.003	-.118	-.089	.613	-.076	.110	.109	.103	.241	-.084	.126
It_cl2_2	.095	-.096	.194	.433	-.054	.316	-.135	.277	-.426	-.007	-.540	-.278	.049	.021
It_cl2_3	-.151	.043	-.368	-.461	.137	-.265	-.323	-.237	.365	-.068	.483	.121	.008	-.109
No_str_1	-.003	.112	-.077	.067	-.068	-.314	.143	-.301	.048	.155	-.102	.313	.217	-.196
No_str_2	.345	.148	-.181	-.137	.142	-.227	-.206	-.182	.288	.196	.114	.268	.190	.113
Age1_2	-.051	.042	.013	.084	-.203	.070	-.010	.059	-.266	-.107	-.052	.017	-.154	.074
Age1_3	-.082	.085	.030	.054	-.097	.013	.073	-.017	-.308	-.063	-.052	.108	-.381	.082
Age2_3	-.075	.087	.033	-.005	.053	-.050	.122	-.088	-.207	.010	-.028	.151	-.425	.051
Alpha2	.106	.059	.145	-.229	.514	-.182	-.179	-.194	.112	-.086	.031	-.070	.456	-.230
Gendiff	.267	-.114	.063	-.042	-.173	.014	-.039	.009	.010	.052	-.272	-.253	.105	-.116
Inv_p	-.119	-.223	-.165	-.338	.376	-.211	-.209	-.193	.516	.086	.389	-.009	.463	-.227
It_tot_c	.067	.056	.131	-.240	.537	-.236	-.211	-.252	.131	-.107	.032	.031	.445	-.231
Itemattr	.243	.183	.114	.489	-.438	.270	.264	.252	-.417	-.002	-.465	.048	-.413	.301
Oplmcla	.009	.019	.074	-.301	.580	-.254	-.244	-.260	.202	-.094	.117	.020	.488	-.249
Oplmdiff	-.273	-.307	-.068	-.274	.516	-.118	-.073	-.125	.301	.003	.233	.172	.400	-.238
Oplmse	.319	.265	.016	.220	-.499	.029	.039	.044	-.087	.004	-.115	-.164	-.347	.198
Sumwc1_1	-.070	-.133	-.134	-.302	.217	-.145	-.110	-.128	.551	.253	.318	.184	.302	-.111
Sumwc1_2	-.122	-.151	.041	-.138	.339	-.180	-.142	-.177	.077	-.008	.086	-.016	.347	-.202
Sumwc1_3	.023	-.030	-.231	-.162	.003	.004	-.079	.034	.319	-.172	.300	-.333	.044	-.039
Sumwc2_1	.122	.251	-.056	-.066	.193	-.071	.230	-.076	.555	.004	.230	.313	.168	.069
Sumwc2_2	-.016	-.161	.157	.134	.107	-.084	-.102	-.094	.109	-.256	-.121	-.404	.219	-.201
Sumwc2_3	-.072	-.145	-.128	-.291	.039	-.160	-.185	-.151	.056	.079	.345	.125	.215	-.228
Tiagrafc	-.093	-.050	.216	-.341	.461	-.406	.047	-.474	.095	-.342	.323	.013	.322	-.390
Tiagrafd	-.152	-.152	.155	-.436	.384	-.160	-.069	-.167	.061	-.015	.334	.132	.354	-.223

A priori item features

	It_frm	it_mng	it_nul	it_cl2_1	it_cl2_2	it_cl2_3	no_str_1	no_str_2
Abs_con	-.022	-.083	.130	.077	.095	-.151	-.003	.345
Base_dis	-.014	-.173	.236	.080	-.096	.043	.112	.148
Basic	-.211	.176	.003	.243	.194	-.368	-.077	-.181
Context	-.175	.276	-.157	.003	.433	-.461	.067	-.137
Distnc	.213	-.282	.132	-.118	-.054	.137	-.068	.142
Freq_bnc	-.211	.386	-.246	-.089	.316	-.265	-.314	-.227
Freq_fin	-.224	.171	.024	.613	-.135	-.323	.143	-.206
Freq_lob	-.231	.359	-.202	-.076	.277	-.237	-.301	-.182
Frq_cc	-.189	-.214	.465	.110	-.426	.365	.048	.288
Imgrated	.179	-.243	.118	.109	-.007	-.068	.155	.196
Imprtnc	-.042	-.182	.278	.103	-.540	.483	-.102	.114
Level	.333	-.293	-.007	.241	-.278	.121	.313	.268
No_hom	.175	-.305	.200	-.084	.049	.008	.217	.190
No_let	.228	-.060	-.169	.126	.021	-.109	-.196	.113
It_frm	1.000	-.651	-.246	-.153	.114	-.011	.025	.268
It_mng	-.651	1.000	-.575	-.119	.200	-.123	-.306	-.508
It_nul	-.246	-.575	1.000	.316	-.378	.170	.363	.361
It_cl2_1	-.153	-.119	.316	1.000	-.376	-.307	-.127	.262
It_cl2_2	.114	.200	-.378	-.376	1.000	-.767	-.093	-.323
It_cl2_3	-.011	-.123	.170	-.307	-.767	1.000	.183	.151
No_str_1	.025	-.306	.363	-.127	-.093	.183	1.000	.114
No_str_2	.268	-.508	.361	.262	-.323	.151	.114	1.000
Age1_2	.130	-.021	-.113	-.108	-.023	.098	-.207	-.029
Age1_3	.085	-.080	.011	.033	-.028	.006	-.077	-.101
Age2_3	.000	-.100	.127	.156	-.020	-.087	.087	-.125
Alpha2	.089	-.265	.242	-.004	.016	-.013	.347	.179
Gendiff	.061	.138	-.241	-.103	.129	-.061	-.186	.074
Inv_p	.044	-.371	.427	-.193	-.135	.272	.336	.195
It_tot_c	.107	-.252	.206	-.073	-.002	.053	.405	.146
Itemattr	-.023	.319	-.382	.180	.206	-.336	-.403	-.155
Oplmcla	.120	-.299	.253	-.123	-.010	.095	.437	.153
Oplmdiff	.088	-.156	.105	-.266	.078	.104	.400	-.071
Oplmse	-.132	.035	.097	.289	-.216	.022	-.342	.164
Sumwc1_1	-.140	-.338	.583	.025	-.241	.229	.255	.194
Sumwc1_2	.471	-.540	.182	-.235	.020	.142	.237	.095
Sumwc1_3	-.484	.527	-.151	-.086	.009	.051	.013	.016
Sumwc2_1	-.192	-.054	.276	.383	-.115	-.147	.066	.376
Sumwc2_2	-.008	-.031	.048	-.249	.443	-.283	-.018	-.041
Sumwc2_3	.059	-.179	.165	-.233	-.391	.564	.251	.185
Tiagrafc	-.011	-.154	.200	.170	-.221	.105	.115	.016
Tiagrafd	.051	-.235	.244	-.013	.001	.008	.272	.010

A posteriori item features

	age1_2	age1_3	age2_3	alpha2	gendiff	inv_p	it_tot_c	itemattr	oplmcla	oplmdiff	oplmse	sumwc1_1	sumwc1_2	sumwc1_3	sumwc2_1	sumwc2_2	sumwc2_3	tiagrafc	tiagrafd
Abs_con	-.051	-.082	-.075	.106	.267	-.119	.067	.243	.009	-.273	.319	-.070	-.122	.023	.122	-.016	-.072	-.093	-.152
Base_dis	.042	.085	.087	.059	-.114	-.223	.056	.183	.019	-.307	.265	-.133	-.151	-.030	.251	-.161	-.145	-.050	-.152
Basic	.013	.030	.033	.145	.063	-.165	.131	.114	.074	-.068	.016	-.134	.041	-.231	-.056	.157	-.128	.216	.155
Context	.084	.054	-.005	-.229	-.042	-.338	-.240	.489	-.301	-.274	.220	-.302	-.138	-.162	-.066	.134	-.291	-.341	-.436
Distnc	-.203	-.097	.053	.514	-.173	.376	.537	-.438	.580	.516	-.499	.217	.339	.003	.193	.107	.039	.461	.384
Freq_bnc	.070	.013	-.050	-.182	.014	-.211	-.236	.270	-.254	-.118	.029	-.145	-.180	.004	-.071	-.084	-.160	-.406	-.160
Freq_fin	-.010	.073	.122	-.179	-.039	-.209	-.211	.264	-.244	-.073	.039	-.110	-.142	-.079	.230	-.102	-.185	.047	-.069
Freq_lob	.059	-.017	-.088	-.194	.009	-.193	-.252	.252	-.260	-.125	.044	-.128	-.177	.034	-.076	-.094	-.151	-.474	-.167
Frq_cc	-.266	-.308	-.207	.112	.010	.516	.131	-.417	.202	.301	-.087	.551	.077	.319	.555	.109	.056	.095	.061
Imgrated	-.107	-.063	.010	-.086	.052	.086	-.107	-.002	-.094	.003	.004	.253	-.008	-.172	.004	-.256	.079	-.342	-.015
Imprtnc	-.052	-.052	-.028	.031	-.272	.389	.032	-.465	.117	.233	-.115	.318	.086	.300	.230	-.121	.345	.323	.334
Level	.017	.108	.151	-.070	-.253	-.009	.031	.048	.020	.172	-.164	.184	-.016	-.333	.313	-.404	.125	.013	.132
No_hom	-.154	-.381	-.425	.456	.105	.463	.445	-.413	.488	.400	-.347	.302	.347	.044	.168	.219	.215	.322	.354
No_let	.074	.082	.051	-.230	-.116	-.227	-.231	.301	-.249	-.238	.198	-.111	-.202	-.039	.069	-.201	-.228	-.390	-.223
It_frm	.130	.085	.000	.089	.061	.044	.107	-.023	.120	.088	-.132	-.140	.471	-.484	-.192	-.008	.059	-.011	.051
It_mng	-.021	-.080	-.100	-.265	.138	-.371	-.252	.319	-.299	-.156	.035	-.338	-.540	.527	-.054	-.031	-.179	-.154	-.235
It_nul	-.113	.011	.127	.242	-.241	.427	.206	-.382	.253	.105	.097	.583	.182	-.151	.276	.048	.165	.200	.244
It_cl2_1	-.108	.033	.156	-.004	-.103	-.193	-.073	.180	-.123	-.266	.289	.025	-.235	-.086	.383	-.249	-.233	.170	-.013
It_cl2_2	-.023	-.028	-.020	.016	.129	-.135	-.002	.206	-.010	.078	-.216	-.241	.020	.009	-.115	.443	-.391	-.221	.001
It_cl2_3	.098	.006	-.087	-.013	-.061	.272	.053	-.336	.095	.104	.022	.229	.142	.051	-.147	-.283	.564	.105	.008
No_str_1	-.207	-.077	.087	.347	-.186	.336	.405	-.403	.437	.400	-.342	.255	.237	.013	.066	-.018	.251	.115	.272
No_str_2	-.029	-.101	-.125	.179	.074	.195	.146	-.155	.153	-.071	.164	.194	.095	.016	.376	-.041	.185	.016	.010
Age1_2	1.000	.758	.165	-.316	.177	.038	-.317	.083	-.288	-.146	.183	-.065	.232	-.214	-.139	-.010	.185	-.075	-.142
Age1_3	.758	1.000	.768	-.226	-.071	-.088	-.220	.091	-.205	-.162	.160	-.108	.127	-.255	-.176	-.174	.066	-.022	-.026
Age2_3	.165	.768	1.000	-.032	-.281	-.170	-.022	.057	-.026	-.101	.062	-.099	-.036	-.176	-.129	-.254	-.081	.041	.099
Alpha2	-.316	-.226	-.032	1.000	-.273	.336	.977	-.413	.961	.546	-.586	.199	.267	.041	.236	.262	.146	.467	.549
Gendiff	.177	-.071	-.281	-.273	1.000	.043	-.299	.084	-.285	-.157	.203	-.171	.117	.132	-.141	.196	-.031	-.185	-.278
Inv_p	.038	-.088	-.170	.336	.043	1.000	.353	-.853	.494	.775	-.518	.727	.627	.188	.321	.426	.557	.330	.494
It_tot_c	-.317	-.220	-.022	.977	-.299	.353	1.000	-.430	.984	.603	-.624	.225	.287	.009	.212	.237	.168	.505	.564
Itemattr	.083	.091	.057	-.413	.084	-.853	-.430	1.000	-.560	-.725	.554	-.609	-.510	-.235	-.169	-.248	-.602	-.425	-.585
Oplmcla	-.288	-.205	-.026	.961	-.285	.494	.984	-.560	1.000	.707	-.701	.315	.380	.046	.229	.275	.262	.530	.627
Oplmdiff	-.146	-.162	-.101	.546	-.157	.775	.603	-.725	.707	1.000	-.920	.541	.511	.142	.226	.340	.411	.455	.637
Oplmse	.183	.160	.062	-.586	.203	-.518	-.624	.554	-.701	-.920	1.000	-.309	-.363	-.139	-.150	-.230	-.281	-.386	-.580
Sumwc1_1	-.065	-.108	-.099	.199	-.171	.727	.225	-.609	.315	.541	-.309	1.000	.106	.031	.355	.069	.331	.246	.397
Sumwc1_2	.232	.127	-.036	.267	.117	.627	.287	-.510	.380	.511	-.363	.106	1.000	-.370	-.117	.522	.378	.205	.295
Sumwc1_3	-.214	-.255	-.176	.041	.132	.188	.009	-.235	.046	.142	-.139	.031	-.370	1.000	.422	.005	.160	.063	.092
Sumwc2_1	-.139	-.176	-.129	.236	-.141	.321	.212	-.169	.229	.226	-.150	.355	-.117	.422	1.000	.162	-.119	.150	.037
Sumwc2_2	-.010	-.174	-.254	.262	.196	.426	.237	-.248	.275	.340	-.230	.069	.522	.005	.162	1.000	-.065	.099	.011
Sumwc2_3	.185	.066	-.081	.146	-.031	.557	.168	-.602	.262	.411	-.281	.331	.378	.160	-.119	-.065	1.000	.261	.424
Tiagrafc	-.075	-.022	.041	.467	-.185	.330	.505	-.425	.530	.455	-.386	.246	.205	.063	.150	.099	.261	1.000	.566
Tiagrafd	-.142	-.026	.099	.549	-.278	.494	.564	-.585	.627	.637	-.580	.397	.295	.092	.037	.011	.424	.566	1.000

Appendix 10: Results of the discriminant analysis.

Variables used in different steps of the discriminant analysis.

Variable	A	B	C	D
abs_con	x	x	x	
base_dis	x	x	x	
basic	x	x	x	
content	x	x	x	
context	x	x	x	
distnc	x	x	x	
freq_BNC	x	x	x	
freq_Fin	x	x	x	
freq_LOB	x	x	x	
freq_rank	x	x	x	
imgrated	x	x	x	
imprtnc	x	x	x	
level	x	x	x	
no_hom	x	x	x	
no_let	x	x	x	
#CLAS1_1	x		x	x
#CLAS1_2	x		x	x
#CLAS1_3	x		x	x
#CLAS2_1	x		x	x
#CLAS2_2	x		x	x
#CLAS2_3	x		x	x
no_CLAS1	x		x	x
no_CLAS2	x		x	x
age1_2	x			x
age1_3	x			x
age2_3	x			x
alpha	x			x
gendiff	x			x
inv_p	x			x
itTotSPS	x			x
itemattr	x			x
itToTOPL	x			x
OPLMDiff	x			x
OPLMSE	x			x
#WCLAS1_1	x			x
#WCLAS1_2	x			x
#WCLAS1_3	x			x
#WCLAS2_1	x			x
#WCLAS2_2	x			x
#WCLAS2_3	x			x
tiagrafc	x			x
tiagrad	x			x

WILKS' LAMDA FOR FIVE CLUSTER CONFIGURATIONS

All variables, one dimension

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	59.111	81.0	81.0	.992
2	7.401	10.1	91.2	.939
3	2.968	4.1	95.2	.865
4	2.616	3.6	98.8	.851
5	.853	1.2	100.0	.679

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 5	.000	109.309	95	.150
2 through 5	.004	62.203	72	.788
3 through 5	.038	37.726	51	.917
4 through 5	.149	21.875	32	.911
5	.540	7.095	15	.955

All variables, two dimensions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	736.580	79.5	79.5	.999
2	158.413	17.1	96.6	.997
3	17.397	1.9	98.5	.972
4	11.649	1.3	99.8	.960
5	1.430	.2	99.9	.767
6	.799	.1	100.0	.666

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 6	.000	223.197	102	.000
2 through 6	.000	143.956	80	.000
3 through 6	.001	83.098	60	.026
4 through 6	.018	48.152	42	.238
5 through 6	.229	17.701	26	.886
6	.556	7.046	12	.855

All variables, three dimensions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	246.791	91.9	91.9	.998
2	10.621	4.0	95.9	.956
3	6.200	2.3	98.2	.928
4	3.601	1.3	99.5	.885
5	1.116	.4	99.9	.726
6	.199	.1	100.0	.407

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 6	.000	142.564	108	.014
2 through 6	.001	79.170	85	.658
3 through 6	.012	50.963	64	.881
4 through 6	.086	28.260	45	.976
5 through 6	.394	10.708	28	.999
6	.834	2.087	13	1.000

All variables, four dimensions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2510.878	99.6	99.6	1.000
2	10.078	.4	100.0	.954
3	.424	.0	100.0	.546

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.000	121.753	63	.000
2 through 3	.063	31.721	40	.822
3	.702	4.064	19	1.000

All variables, five dimensions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	75.909	64.1	64.1	.993
2	22.772	19.2	83.4	.979
3	19.679	16.6	100.0	.976

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.000	121.213	63	.000
2 through 3	.002	71.273	40	.002
3	.048	34.835	19	.015

STRUCTURE MATRICES FOR CLUSTER GROUPINGS ONE DIMENSION, ALL VARIABLES

	Function				
	1	2	3	4	5
it_cl1_1	.431	-.098	-.211	.026	-.173
it_cl1_2	-.411	.090	.037	-.011	.024
it_cl2_3	-.377	-.052	.206	-.331	.037
no_str_2	.344	-.063	.273	-.274	.075
sumwc1_3	-.337	.103	-.146	.275	-.098
no_lett	.307	-.082	-.126	-.004	.253
no_hom	-.219	.014	.019	.193	-.024
no_str_1	.006	.403	.000	.026	-.113
tiagrafd	.023	.358	-.150	-.057	.161
OPLMdiff	-.052	.351	-.220	-.027	-.017
OPLMSE	.089	-.334	.322	-.029	.026
itTotOPL	-.081	.283	-.022	.097	.027
It_tot_c	-.088	.278	.034	.105	.044
itemattr	.017	-.219	-.070	-.105	-.009
freq_LOB	-.061	-.194	.018	-.057	-.074
tiagrafc	.074	.183	.018	-.065	.016
sumwc1_2	.119	.178	.099	-.064	-.057
freq_BNC	-.054	-.177	.014	-.054	-.066
alpha	-.024	.165	.070	.121	.056
context	-.030	-.162	.098	-.133	.057
inv_p	.002	.088	-.033	-.072	-.055
it_cl2_2	.190	.173	-.332	.270	-.291
it_cl1_3	-.022	.008	.199	-.017	.171
sumwc1_1	-.059	-.081	-.181	-.156	.004
distnc	.035	.032	.079	-.012	-.053
level	-.052	.219	-.156	-.250	.115
sumwc2_3	-.084	-.009	-.178	-.208	.161
imprtnc	-.080	-.199	-.040	.200	.183
gendiff	.006	-.057	.138	-.182	-.024
freq FIN	.041	-.050	.011	-.099	.011
sumwc2_2	.124	.153	.131	.074	-.311
it_cl2_1	.154	-.151	.183	.015	.303
freq rank	-.012	.129	.005	.061	.252
age1_2	.046	-.026	.041	-.047	-.192
base_dis	.008	.067	-.048	.136	-.187
sumwc2_1	-.170	-.068	.072	.119	.184
gendiff (absolute)	-.033	.146	-.056	.101	.182
abs_con	.011	-.069	.022	-.137	.165
imageab	.020	-.101	-.102	-.143	.164
basic	-.049	-.031	.056	-.076	.134
age2_3	.008	.000	-.048	-.005	.072
age1_3	.033	-.015	-.006	-.032	-.070

TWO DIMENSIONS, ALL VARIABLES

	Function					
	1	2	3	4	5	6
tiagrafd	.525	.240	.037	.119	-.206	-.236
it_cl2_2	.437	-.293	.313	-.083	-.125	-.164
no_str_2	-.421	.149	.223	-.107	.114	-.230
OPLMSE	-.392	.136	.072	-.327	.131	-.357
OPLMdiff	.379	-.046	-.084	.270	-.081	.335
no_hom	.351	-.141	-.099	.076	.057	.047
sumwc2_2	.305	-.058	.052	-.257	-.134	-.129
it_cl1_1	-.077	-.442	.313	.007	.037	-.115
imageab	-.091	-.255	.025	-.173	-.154	-.084
no_str_1	.058	.181	-.015	-.079	.023	-.013
level	-.058	-.166	-.164	-.086	-.058	-.066
age1_3	-.004	.103	-.085	-.011	.049	.024
age1_2	.012	.063	.038	-.025	-.057	.038
it_cl2_3	-.370	.221	-.418	.155	-.022	.398
age2_3	-.021	.087	-.185	.015	.151	-.009
alpha	.022	-.025	-.305	.530	.143	-.090
itTotOPL	.137	-.023	-.352	.524	.181	.077
itTotSPS	.044	-.029	-.393	.490	.169	-.006
distnc	.007	-.001	-.116	.235	.159	.175
basic	.011	-.011	-.057	.169	-.115	-.012
itemattr	-.259	-.374	-.049	-.237	-.425	-.233
tiagrafc	.108	.284	.060	.246	.357	.145
gendiff	.027	.011	.181	-.044	-.268	.055
sumwc2_1	.071	.003	-.179	-.121	.260	.021
base_dis	-.012	-.007	-.016	-.076	.250	-.165
gendiff (absosulte)	.024	-.052	-.048	.191	.235	-.065
freq_rank	-.001	-.019	-.114	.120	.139	-.057
freq_BNC	-.006	.001	.032	-.045	-.076	.009
freq_LOB	-.006	.001	.030	-.046	-.073	.007
abs_con	-.005	.032	.131	-.002	-.012	-.500
sumwc1_3	.408	.448	.145	.205	.140	.450
inv_p	.349	.154	.174	.039	.136	.446
sumwc1_2	.203	-.243	-.125	-.153	-.052	-.413
sumwc2_3	-.057	.148	.112	.274	.151	.406
it_cl1_2	.108	.204	-.334	.219	-.167	.401
sumwc1_1	-.101	.033	.112	.007	.077	.297
it_cl1_3	-.028	.281	.003	-.241	.136	-.296
context	-.008	.012	.131	-.145	-.131	-.269
no_let	-.152	-.021	.186	.047	-.190	-.268
it_cl2_1	-.148	.131	.075	-.072	.193	-.250
freq_FIN	-.004	-.001	.031	-.065	-.091	-.134
imprtnc	-.001	-.013	-.038	.048	.099	.129