– LANCASTER UNIVERSITY –

DEPARTMENT OF LINGUISTICS & ENGLISH LANGUAGE

ARE THREE OPTIONS BETTER THAN FOUR?

Investigating the effects of reducing the number of options per item on the quality of a multiple-choice reading test

GERARD SEINHORST

Dissertation submitted in partial fulfilment of the requirements for the M.A. degree in Language Testing (by distance)

December 2008

19,997 words

ABSTRACT

The present study investigated the effects of reducing the number of options per item from four to three on the psychometric characteristics and completion time of a multiple-choice reading test. Statistical analyses showed that the effects on mean item difficulty, mean item discrimination and internal consistency reliability were nonsignificant. These results are consistent with most previous research. Distractor analyses revealed that most likely the limited effectiveness of many distractors may have accounted for the nonsignificant findings: only 17% of the 4-option items had 3 effectively functioning distractors, rendering the 4-option test essentially a 3-option test for the majority of the test takers.

On average, the 3-option items were completed approximately 9% faster than their 4-option counterparts, thereby increasing the efficiency with which information on test taker ability is gathered.

This study demonstrated further that, as a group, subject matter experts exhibited fairly high ability to detect without statistical data which distractors of 4-option items will be chosen least frequently by test takers. This suggests that it must be possible to develop a 3-option test from the beginning that is equally reliable and discriminating as a 3-option test created based on 4-option item statistics.

Finally, this study extended and further supported the practical advantages of 3-option test items by presenting evidence that more than 40% of the test takers preferred questions with 3 options, while only 7% favoured 4 answer choices per item. Irrespective of their ability level, students generally perceived the 3-option format as more efficient, less confusing and equally acceptable as the traditional 4-option format.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Judit Kormos for her invaluable comments, suggestions, and encouragement during the writing of this dissertation. Also, I would like to thank my tutors at Lancaster University: Professor J. Charles Alderson, Dianne Wall and Jayanti Banerjee for giving me many new insights in the field of language testing. I am very grateful to Dr. Martha Herzog for her useful advice and to everyone who gave up their precious time to participate in this study. And last but not least, I owe a big thank you to my wife Silvia who has supported me throughout my studies, and of course to my new-born daughter Irene who permitted me to work undisturbed for many nights.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

The beginning of knowledge is the discovery
of something we do not understand.
FRANK HERBERT (1920 - 1986)

One of the most widely used test formats in language testing is the multiple-choice (MC) test, favoured for a number of reasons: MC tests provide better coverage of the content and the processes to be assessed than many other test formats; MC items can be easily pretested, used and reused; objective test scoring is possible, thereby leading to increased reliability; and most kinds of content can be tested using this format, including many types of higher level thinking (Haladyna & Downing, 1989: 37-38). Ebel (1972: 187-188) maintains that MC items are likely to be less indirect and artificial than some other question types, that test takers often find MC questions less ambiguous than completion or true-false items, and that teachers usually find it easier to defend the correct answer to their students. According to Hopkins (1998) MC items may generate a desirable ambiguity extrinsic to the item itself (as opposed to the undesirable intrinsic ambiguity) as a result of faulty understanding on the part of the test taker. The major virtue of MC tests is then that they can 'require the test taker to *discriminate* among alternatives that can require a level of mastery that a free-response item may not be able to detect' (p. 214).

MC test items take many forms, but in their basic structure they consist of at least the following parts: (a) a stem: the stimulus for the response; (b) the correct choice or "key"; and (c) several wrong answers, usually referred to as "distractors". The effectiveness of MC questions depends on the validity of the stem and the key, but we should not lose sight of the importance of the distractors in making MC questions efficient.

Distractors are often considered the most difficult part of the MC test item to write. A distractor is an unquestionably wrong answer, but it must be plausible enough to attract test takers who have not yet learned the knowledge or skill that the test item is supposed to measure. To those who possess the knowledge asked for in the item, the distractors are clearly wrong choices. Each distractor should resemble the correct choice in grammatical form, style, and length. At the same time, distractors should not overlap, subsume, or be synonymous with one another. In order to be effective, a distractor should have a degree of discrimination, that is, should be appealing to low-ability students and be rejected by high-ability students. If no one chooses a particular distractor, it is not participating effectively in the process of giving the question a factor of difficulty. On the other hand, a distractor which is too attractive may be a correct answer to a badly posed question.

One issue that must be faced whenever a new MC test is constructed is the number of answer choices to include per item. Typically, 4 or 5 alternatives are used for each item, usually based on the rationale that it improves the psychometric quality by limiting the effects of random guessing. Random guessing, if present, will tend to lead to the overestimation of achievement. At the same time, as Haladyna (2004) states, 'item writers are often frustrated in finding a useful fourth or fifth option because they typically do not exist' (p. 112). Alderson *et al.* (1995) recommend using 4 options, but they also add that 'if it is impossible to think of a third attractive wrong answer, then it is sensible to have only three alternatives for some items' (p. 48).

In addition to streamlining the item writing process, using only 3 options would have another distinct advantage: it will shorten the time taken to complete the test by reducing the amount of material that a test taker needs to read and analyse. This suggests that using 3-option items has benefits in terms of efficiency, as the total testing time can be reduced. Alternatively, more items could be administered without an increase in time, which may result in a more content valid test.

The question arises, then, whether a 3-option test can be as reliable and valid as a 4-option test, rendering it a viable alternative to the traditional 4-option format. Although research on educational and achievement tests suggests that MC tests containing items with 3 options frequently have reliabilities equal to or greater than tests composed of items containing 4 or 5 options, the topic of the optimum number of alternatives continues to be a matter of considerable controversy, which is apparent, for instance, from the recurring debates on this topic at LTEST-L, an online discussion forum dealing with language testing issues.

The present study addressed this question by investigating the possible differences in test scores and item performance between a reading test consisting of 4-option items and a parallel test with 3-option items. The intent was to investigate whether empirical findings from previous studies, using primarily tests of rote knowledge, are generalisable to a test that measures reading comprehension in a foreign language.

In addition, in this study differences in the completion time of both tests were systematically monitored, in order to explore whether the assumed practical benefit of reducing the total testing time using 3-option items also counts for tests of reading comprehension.

Further, the study sought to determine how reliably non-functioning distractors can be identified intuitively, that is, without any empirical item analysis data. Research by Alderson (1993) suggests that expert judges are often unable to agree upon the difficulty level of a particular reading comprehension item, and it would be valuable to find out whether they are better able to predict the efficiency of a particular distractor. The trustworthiness of such intuition would greatly help to optimise the item-writing process.

A final aim of the study was to evaluate test takers' perceptions and attitudes toward the 3-option format in comparison with both the traditional 4-option format and the candidates' actual test performance. If could be demonstrated that test takers perceive the 3-option format as an acceptable alternative to the 4-option format, this would add to the theoretical and statistical evidence favouring the use of the 3-option test item.

The results of this study may have useful implications for test administrators and test writers using MC items. In many large-scale testing programs the press for more efficient and reliable measurement is often a major concern of measurement specialists and test administrators. The results of this study may help testing programs considering moving to the 3-option format to make a better informed decision. This study may also provide valuable information in the context of testing in the classroom or in training programs, where the objective of a MC test is to obtain a measure of student learning in an efficient manner. Writing good MC items is a time-consuming task, especially for non-professional item writers, so if it is possible to make the item-writing process more efficient without sacrificing the test quality, it would be a benefit to all concerned. Information that increases our understanding of MC items and tests will improve our ability to measure student achievement and other constructs. Improved information will lead to improved item writing, improved test design, better measures of achievement and skill level, and more appropriate score interpretation and decision making.

This dissertation is organized as follows. Chapter 2 presents a synthesis of the most relevant literature on this topic, providing the theoretical context of the study and the justification for the research questions. Chapter 3 outlines the methods used for the collection and analyses of the data, and the rationale behind these methods. Chapter 4 provides a detailed account of the results from the analyses of the collected quantitative and qualitative data, followed by a discussion of the answers to the research questions in Chapter 5. In the last chapter the main findings are summarized, and some recommendations for further research are given.

CHAPTER 2: THEORETICAL CONTEXT

It is not necessary to understand things
in order to argue about them.
PIERRE BEAUMARCHAIS (1732 - 1799)

2.1 BACKGROUND

The optimal number of options for a MC test item has always been an issue of considerable interest to testing specialists. In a review of 46 textbooks covering the topic of writing multiple-choice test items, Haladyna and Downing (1989: 45) reported that authors often disagree on the ideal number of options. Most recommend writing as many plausible distractors as possible, and a few textbook authors emphasize 4 options, which according to Haladyna *et al.* (2002: 317) may have become the standard MC item format in the testing industry. Three-option test items are not typically recommended for testing programs, because theoretically fewer options will lower the test reliability and increase the possibility that poor students raise their score by guessing. On the other hand, creating fewer distractors and administering items which take up less space and require less reading and processing by the test takers will most likely reduce test development and applications costs, and shorten test administration time. A 3-option test would thus have advantages over 4- and 5-option tests with respect to another important test criterion: practicality.

In light of the potential benefits of reducing the number of options, over the past decades a great number of studies have examined the effects of changing the number of options on the psychometric properties of a MC test. The first studies investigated the issue mainly from a theoretical perspective.

2.2 THEORETICAL STUDIES

2.2.1 *Effects on test quality*

In 1964 Tversky presented mathematical proof that the "power" of a MC test (defined as one minus the probability of getting a perfect score by chance alone), its ability to discriminate between test takers, and the amount of information a test can provide are factors that are all optimised in a 3-option test compared to other types of MC tests when the total number of options is fixed. An example of satisfying Tversky's criterion would be in the comparison of a 30-item 4-option test with a 40-item 3-option test; the total number of options is fixed, as both tests contain 120 options or choice points.

In another theoretical paper Ebel (1969) predicted, according to a variant of the Kuder-Richardson Formula-21 he developed, that a considerable increase in the reliability of a 100-item objective test will occur when the number of options is increased from 2 (expected *r* =.74) to 3 (.84), a smaller increase when 4-option items (.86) are used, and a still smaller increase beyond that point (p. 565). This suggests that, assuming that it takes less time to write and answer a 3-option item compared to a 4- or 5-option item, 3-option items would represent the best compromise between maximum test reliability and efficiency.

Grier (1975) extended Ebel's formula to estimate optimal reliability and concluded that if the number of options in the total test is held constant, 3-option items give a test that is theoretically 'more reliable, more powerful, more discriminative, and more informative' (p. 112). He found that this superiority held provided there were more than 54 alternatives (i.e., more than 18 3-option items) in the test. Grier (1976) later advanced Tversky's (1964) argument by generalizing the goal of optimising the number of alternatives under a fixed total time. He demonstrated how allowing for "travel time", that is, the time it takes to read a question and to consider each option (where Tversky considered no time between items and time as a linear function with the number of options) yielded Tversky's optimum result exactly.

2.2.2 *Effects on student performance*

Lord (1977) has reviewed the Tversky (1964) and Grier (1975) approaches and explored theoretically, under Tversky's condition of a fixed total number of alternatives, two further approaches. Under the assumption that all wrong answers are guessed wrong and that all correct answers are obtained either by knowledge or by random guessing, and using test reliability as the criterion he found 3-option items to be optimal if they were hard items or of moderate difficulty. Lord's fourth approach is based upon item characteristic curve theory. Using item parameters from an actual test administration he simulated tests composed of 2-, 3-, 4- and 5-option items of moderate difficulty holding the total number of options, item difficulty and discriminating power constant across test forms. He found the test composed of 3-option items to be superior to the 2-, 4- and

5-option tests, but added that reducing the number of options results in a more efficient test for high-level students and a less efficient test for low-level students. At the lower range 4- and 5-option items work best, because ‘at low ability levels the effect of random guessing becomes of overwhelming importance’ (p. 36), and guessing erodes validity especially when items have fewer options.

2.2.3 *Recent approaches*

Bruno and Dirkzwager (1995) investigated the optimal number of options for MC items through an information-theoretic perspective. They argued that information from the test item as a whole generally increases with the number of options, but the mean information content per option on a test has a maximum point. ‘Conceptually, too many alternatives to a multiple-choice test item introduce noise into the test item, which results in little or diminishing marginal information being extracted’ (p. 962). They demonstrated that maximum information was obtained on test items with 3 options under the condition where each option had an equal probability of being answered (equally plausible) by an uninformed individual.

More recently, Abad *et al.* (2001) manipulated a 5-option vocabulary test in order to create 2-, 3-, and 4-option test forms. The answers of 452 test takers to the worst alternatives were randomly reassigned to generate their hypothetical answers to items with fewer options. Using a procedure based on Item Response Theory, changes in item parameters, test information function and ability estimation were analysed. The (hypothetical) results on the 3- and 4-option tests hardly differed from those obtained in the original 5-option format. The authors concluded that dropping one or two options will not seriously harm MC items.

2.2.4 *Underlying assumptions*

Although the theoretical evidence consistently suggests that the 3-option format is optimal, the validity of some of the assumptions on which the findings are based can be challenged. In the first place, these studies have had the assumption of knowledge or random guessing when a test taker is confronted with an item. Not only does this suppose that test takers do not have partial knowledge about item content – which is rather improbable in most educational testing situations –, but also does this assumption contradict more recent findings that MC items may be susceptible to testwiseness (Rogers & Yang, 1996).

Another assumption common to most theoretical solutions is that all distractors are equally attractive. According to this assumption, tests with different numbers of options can be obtained either by arbitrarily eliminating distractors or by randomly sampling options to be eliminated. In reality, as demonstrated by Haladyna and Downing (1993), the various options are not equally attractive, and therefore reducing the number of options may in actual practice lead to different outcomes from what can be predicted on the basis of mathematical formulas.

A last assumption held in these theoretical studies is the law of proportionality. This law states that the time needed to respond to each item is a function of the number of options. Thus, the more options in each test item, the longer it takes to complete the test. Budescu and Nevo (1985) took issue with the assumption of proportionality. In their empirical study they found a strong negative relationship between rate of performance and the number of options for tests of fixed numbers of items. They argued that testing time depends on the number of items, the number of options, and a function of the item's complexity, making change in response time not a simple function of the number of options.

2.2.5 *Summary*

These theoretical and test simulation studies reveal that, irrespective of the approach taken, 3-option MC tests are at least as good as, and in some cases even superior to 4- and 5-option MC tests in terms of item discrimination at the item level and internal consistency reliability at the test level, although Lord's (1977) study suggests that breakdown of test takers by ability may require further elaboration. However, given the uncertainty about the validity of the assumptions underlying most of the theories, the conclusion suggesting the theoretical optimality of 3-option items remains questionable without empirical support.

2.3 EMPIRICAL STUDIES

Empirical research on the optimal number of options covered a wide range of conditions, subject areas, test taker samples, and methods used. Of primary concern in most studies were the effects of reducing the number of options on the psychometric properties of MC tests. Generally, the changes in the difficulty indices, in the items discrimination indices and in the test reliability were analysed.

2.3.1 *Effects on test quality*

Stimulated by Tversky's (1964) paper, Costin (1970) conducted an empirical study in which he randomly eliminated the fourth option from a pool of psychology achievement items. The two tests consisting of 3- and 4-option items were administered to a sample of students. Costin found that his 3-option items were less difficult, but more discriminating and more reliable than the 4-option items. However, on the basis of Ebel's (1969) theoretical analysis one would have expected a decrease in reliability associated with the random reduction of alternatives in test items.

A second study by Costin (1972) yielded higher reliabilities, estimated via the Spearman-Brown formula, for sets of 3-option items compared with 4-option items under Tversky's condition of a fixed number of choice points, but not under the condition of an equal number of items per set. Costin concluded that the results of this study confirmed the practical benefits of 3-option items without sacrificing the reliability and validity of a 4-option item test: increasing the efficiency of the assessment by a reduction of the test completion time, and a less arduous and time consuming job for the test writer (p. 1037).

Crehan *et al.* (1993) compared 4-option psychology test items with 3-option items where the least discriminating distractor was dropped. In line with their expectations, the 4-option items were found to be significantly more difficult than the 3-option items, but the reduction of the number of options did not affect item discrimination. The authors presented the findings as further evidence of the efficacy of 3-option items.

Most recently, Shizuka *et al.* (2006) investigated the effects of reducing the number of options per item on the psychometric characteristics of an English MC reading test. Responses to the two tests indicated that using 3 options instead of 4 did not significantly change the mean item difficulty or the mean item discrimination. Their findings suggested that 3-option items performed nearly as well as their 4-option counterparts.

The results of several other studies tend to agree with most recurrent outcomes of the studies reviewed above: tests consisting of 3-option items are at least equivalent to 4- or 5-option tests in terms of internal consistency score reliability when the number of choice points in the whole test is fixed (Trevisan *et al.*, 1991; Haladyna & Downing, 1993; Delgado & Prieto, 1994; Sidick *et al.*, 1994; Rogers & Harley, 1999), and difficulty is inversely related to the number of options (Straton & Catts, 1980; Owen & Froman, 1987; Cizek & O'Day, 1994; Trevisan *et al.*, 1994; Berríos *et al.*, 2005). Although in some of the studies significant differences were found, the magnitude of these differences tended to be very small, and with limited practical implications. With regard to item discrimination, the results are less consistent. Some studies reported no changes in discrimination (e.g., Ramos & Stern, 1973; Owen & Froman, 1987; Delgado and Prieto, 1994), others found improved discrimination values for items with fewer options (e.g., Straton & Catts, 1980; Trevisan *et al.*, 1991; Berríos *et al.*, 2005), whereas the findings for mean item discrimination in studies by Budescu and Nevo (1985), Cizek *et al.* (1998), and Rogers and Harley (1999) were not conclusive.

2.3.2 *Effects on student performance*

In his above-mentioned theoretical study Lord (1977) found a relationship between the optimal number of options on MC items and the ability level of the test takers. Lord's findings suggested that decreasing the number of options results in a more efficient test for high-level test takers but in a less efficient test for low-level examinees. Green *et al.* (1982) investigated Lord's argument by comparing the reliabilities and validities of 3-, 4-, and 5-option teacher-made MC tests for low-, average-, and high-ability level students. The results of their study did not support Lord's theoretical predictions regarding test reliability (KR-20) in a classroom situation. For the low-ability group, test reliability was highest for the 4-option test, significantly lower for the 3-option test, and worst for the 5-option test. For the high-ability group, differences among reliabilities were not significant. The optimal number of alternatives for all ability groups combined was found to be four. According to the authors, a likely explanation for the discrepancy with Lord's results was that the tests used in Lord's study were considerably more difficult than their tests, as a result of which the differences in item responses between ability groups might have been reduced. Moreover, the range of student abilities in their study was quite narrow. Instead of contrasting low- and high-ability students on a difficult test as was done by Lord, their study compared moderately high and slightly higher ability students on an easy test. However, the authors argued that the conditions of their study were more representative of the typical classroom test than were those in Lord's study.

A study by Levine and Drasgow (1983) supported Lord's (1977) conclusion that high-ability test takers may be less inclined to guess, thereby not needing as many options as low-scoring students who are more inclined to guess. In examining the relationship of incorrect choices and ability, they found that, with more able students, only one or two distractors were typically selected. As ability level decreased, there was a tendency for more of the distractors to be chosen by the lower ability groups. These results suggest that information is maximized by using more options per item for lower ability groups and fewer numbers of options for higher ability groups.

Finally, Trevisan *et al.* (1991) examined this issue with a sample of secondary school students of varying levels of ability. They used 3-, 4- and 5-option versions of the same test where the least discriminating options were omitted for the 3- and 4-option versions. They found that the reduction of number of options had no substantial effect on internal consistency reliability. Thus, their study did not corroborate the findings that maximum reliability would be obtained for high-ability students as the number of options per item decreases, and that the maximum reliability would be found for low-ability students as the number of options per item increases.

2.3.3 *Effects on test efficiency*

In numerous studies investigating the optimal number of options in a MC test item (Costin, 1972, 1976; Grier, 1976; Haladyna & Downing, 1993; Cizek & O'Day, 1994; Sidick *et al.*, 1994; to cite just a few) it has been implied that using fewer options has distinct advantages in terms of efficiency: test takers would need less time to read and process the options, leading to a reduction of the administration time or, keeping the time constant, in a better sampling of the content as more items can be included in the test. Even if Tversky's (1964) assumption that the duration of test-taking time is proportional to the total number of options in the test does not quite hold (Budescu & Nevo, 1985), there is some empirical evidence that it takes less time to respond to a 3-option item than to a 4- or 5-option equivalent.

For instance, Straton and Catts (1980) systematically observed the average time taken to respond to items varying in number of alternatives from 2 to 4, and found that the mean time per item decreased with the number of alternatives per item. The mean time to complete an item was found to be .56, .77, and .91 minutes for test forms consisting of 2, 3 and 4 options respectively. It should be noted, though, that the reverse appeared to be true for the entire tests when keeping the number of choice points fixed: the mean test completion time for the 60 item 2-option test was 33.62 minutes, for the 40 item 3-option test 30.68 minutes, and for the 30 item 4-option test 27.25 minutes. This indicates that a comparatively large amount of time is spent on reading the item stem, and relatively less time on reading the options.

Rogers and Harley (1999), on the other hand, found that the times required to complete 3- and 4-option forms of a mathematics examination were essentially the same. But this was probably attributable to the observation that the mathematics items required the students to perform time-consuming calculations rather than simply to recall short answers.

Owen and Froman (1987) reduced 5-option items from a psychology examination to 3-option items by discarding the least discriminating distractors. Apart from the fact that they found no substantive difference in the item difficulty or discrimination of identical 3- and 5-option items, they did note a difference in administration time in favour of the 3-option format. The 3-option form took 17% less time which suggests that an additional eight or nine items (keeping testing time constant) would improve both content-related validity and reliability. Test takers themselves seemed to prefer fewer options as well. As part of their study, Owen and Froman asked the 114 participants to vote for their preferred form: 111 (97.4%) voted for the 3-option form, 3 had no preference, and none chose the 5-option form.

Although the empirical results are not unequivocal, most studies support the idea that fewer options decrease the administration time, even the Budescu and Nevo (1985) study, which rejected the law of proportionality. If administration time is shortened by using test items with fewer options, then it appears that one can administer more items in the same period of time and produce more reliable test scores. Another factor, not considered in these studies and relating to efficiency, is the time gained by constructing items with fewer options. Without more empirical evidence, it seems reasonable to conclude that the use of fewer options leads to greater efficiency.

2.3.4 *Main results*

Despite the differences in the research methods used, the results of these empirical studies seem to support the following conclusions:

- the reduction of the number of options in MC test items from 4 to 3 has little psychometric effect; more specifically, using fewer options (a) slightly reduces the difficulty; (b) does not affect in a systematic, practical and/or significant way neither the item discrimination nor the internal consistency reliability;
- using fewer options generally yields a more efficient test in terms of administration time, though it is not clear if a systematic relation exists between the time to complete a test and the number of options; most studies agreed that by administering more items in a given time, the content validity and test reliability could be improved;
- the "optimal" number of options may not be independent of the test takers' levels of ability: there is some inconclusive evidence to suggest that reducing the number of options is to the disadvantage of lower ability test takers;
- 3-option items seem generally to be preferred by the test takers, although few studies actually investigated this aspect.

The unsystematic changes in reliability and discrimination are contrary to expectations from standard measurement theory and require further explanation. Generally, more items – or more precisely: more choice points – on a MC test leads to greater reliability (Weitzman, 1970: 83; Hopkins, 1998: 121). Thus it would seem that the use of more options does increase reliability systematically but, as Ebel (1969) demonstrated in his earlier mentioned study, beyond 3 options per test item the extent of this increase will usually be marginal (in the range of .02 to .05). The reported inconsistencies in the changes in reliability may further be explained by the considerable differences in sample sizes, and above all by the differences in methods used: in some studies the number of choices per test were kept fixed, whereas in other studies only the number of items of the tests were held constant, which effectively shortens a 3-option item test.

There is another factor that might play a role and which is related to the observation that in many cases the discrimination was not significantly affected. An explanation for this finding may be found in distractor analyses. Haladyna and Downing (1993) examined the effectiveness of distractors on a 5-option, high-quality standardized medical education test. A distractor was considered to be functional if it has (a) a significant negative point-biserial correlation with the total test score, (b) a negatively sloping item characteristic curve, and (c) a frequency of response greater than 5% for the total group. They found that items with 2 or 3 functional distractors were very rare (1-8%); the average number of functional distractors per item was about one. Also, the number of effective distractors was unrelated to item difficulty and positively related to item discrimination. Similarly, distractor analyses conducted by Shizuka *et al.* (2006) on a reading comprehension test revealed that regardless of whether 4 or 3 options were provided, the actual test takers' responses spread, on average, over about 2.6 options per item, that the mean number of functioning distractors was much lower than 2, and that reducing the least popular option had only a minimal effect on the performance of the remaining options.

The findings of these studies indicate that no matter whether 3, 4 or 5 options are used, most of the job is being done by two or fewer functioning distractors. This would explain the observation that discrimination is often not much affected by a reduction of the number of options.

2.3.5 *Summary*

Most of the authors of the studies reviewed here concluded by recommending the 3-option format, either because they found no significantly different item performance between the 3-option format and formats with more options or because, even when they did, the effect size was negligible. At the same time, many authors recognized that the issue of the optimal number of options is still a matter of considerable debate, and that its solution demands further evidence; most studies concluded with a plea for more research in different contexts with different samples and different MC tests.

It would be reasonable to state, then, that applying the 3-option format deserves much more serious consideration than is commonly given by writers of standardized tests. As a matter of fact, the only recurring objection to using fewer options is that it adds to the chance of an examinee *guessing* the right answer. Minimizing the presence of guessing is one of the main reasons why a great and reasonable number of options (4 or 5) is traditionally advised. But, as Hopkins (1998: 148) asserts, the effects of guessing are often overrated. The probability of guessing the right answer is a ratio of one to the number of options, which means that, provided 3 options per item, the probability of guessing correctly on ten items is about .0000015. Moreover, as studies by Ebel (1968: 324) and Bussis and Chittenden (1987, cited by Farr *et al.*, 1990: 224) revealed, well-motivated test takers who have time to attempt all items do relatively little blind guessing, and what is usually referred to as guessing is often an educated selection based on partial knowledge. Thus, it seems that, especially in longer tests, the effect of guessing on a MC test is often negligible.

2.4 THE RESEARCH TOPIC

2.4.1 *The research gap*

Although an extensive body of research investigated the effect of the number of options on the psychometric quality of a test, few studies appear to have addressed this issue in the field of foreign language testing. In most studies the tests measured rote knowledge or reproduction skills, and in the rare cases that foreign language testing was the area of investigation, the instruments used were either vocabulary tests or achievement tests. Therefore, it is largely unknown to what extent the findings from previous studies are generalisable to tests that measure general reading or listening comprehension in a foreign language. In such tests the options must somehow, in order to be minimally plausible, be based on the actual item passages rather than on outside knowledge. Conceivably, this prerequisite puts greater constraints on the possibility to write several effective distractors for comprehension tests as compared to knowledge tests. This would be the case in particular at lower language proficiency levels where item passages are necessarily short and the opportunities to construct many plausible distractors are generally limited. The effects of reducing the number of options on the item discrimination for comprehension tests may therefore yield different outcomes than those from previous studies.

Further, Lord (1977) and Levine and Drasgow (1983) found experimental evidence indicating that decreasing the number of options resulted in a more efficient test for high-level examinees but a less efficient test for low-level examinees. However, since these findings were not or were only partially corroborated in studies by Green *et al.* (1982) and Trevisan *et al.* (1991), it appears that this issue merits further investigation.

More research seems also desirable with respect to the effect of reducing the number of options per item on the test completion time. It has been suggested that substantial gains in administration time can be achieved by using fewer option. However, the research findings with regard to test completion time are not conclusive, as Rogers & Harley

(1999) found in their study that the times required to complete 3- and 4-option forms were essentially the same. Review of the empirical literature reveals further that the potential differences in test completion time have not been systematically examined with regard to foreign language tests. Presumably, in a comprehension test the time needed to complete each item is spent mainly on listening to or reading the passage and proportionally less on answering the actual question. Consequently, reducing the number of options per item in this kind of tests may have a less noticeable effect than has been reported in previous studies. In this context, Straton and Catts (1980) remarked 'where item stems necessarily will require long reading times relative to an alternative, say substantially more than twice as long, … then the use of four or even five-choice items would seem to be more desirable' (p. 364). However, to date no empirical support for the validity of this assumption has been reported.

Furthermore, an important methodological issue that has been largely neglected so far relates to the extent to which content specialists are able to judgementally (i.e., without statistical data) identify non-functioning distractors in 4-option items. The supposed savings in development time using 3-option items will be lost if the only way to obtain a 3-option test equally valid and reliable as a 4-option test, is by reducing the fourth option on an empirical basis. Not only would this imply that the "problematic" fourth option always needs to be written anyway, but also that in order to create a reliable 3-option test time-consuming pretesting and item analyses would always be required. However, as Shizuka *et al.* (2006: 41) pointed out, concern over fairness and test security makes it not always possible to pilot test items before the actual administration. Thus, only if evidence can be found that item writers are able to reliably determine *in the actual item writing phase* which distractor would be potentially non-functioning, the 3-option format – provided that the psychometric qualities are indeed the same as the 4-option format – can be regarded as a truly efficient alternative to the 4-option item. Until then, research findings suggesting that from a practical perspective 3 options is the optimum number of choices for an MC test must be deemed questionable.

Finally, given the apparent psychometric and practical advantages of using 3-option MC tests, it is noteworthy that hardly any studies have been undertaken to investigate the attitudes and perceptions of the test takers regarding the 3-option format. In light of the – albeit inconclusive – findings in the earlier mentioned study by Lord (1977), the number of options may influence the performance of high-level test takers differently than that of low-level test takers, and it is not unthinkable that this affects the way test takers with varying ability levels perceive a 3-option test. At any rate, if generally test takers had a negative attitude towards such a test or perceived it as less valid, it might inhibit the face validity, the acceptance, and ultimately the application of this test format, despite its alleged benefits.

2.4.2 *The research questions*

Emerging from the issues raised above, the following research questions were developed:

1. What are the effects of reducing the number of options per item from 4 to 3 – other factors being equal – on the psychometric properties and completion time of a multiple-choice reading test? Specifically, what influence does reducing the number of options have on the performance of test takers with differing proficiency levels?
2. How reliably can test writers intuitively predict which distractor of 4-option items is functioning least well? In other words, to what extent do judgementally deemed non-functioning distractors agree with judgements that would be made based on the actual statistical performance of the items?
3. Which differences, if any, are there in test takers' attitudes and perceptions of a 3-option test compared with a 4-option test? In particular, does decreasing the number of options affect the attitudes and perceptions of high-performing test takers differently than low-performing test takers?

The scope of this study was restricted to finding answers to the above research questions in an attempt to fill the identified research gap. Therefore, broader issues such as the nature of reading in a foreign language, or whether multiple-choice tests are the most appropriate way of assessing reading comprehension were not addressed. Nor did the study serve as a validation of the reading test that was used as one of the instruments for data collection.

CHAPTER 3: METHODOLOGY

Art and science have their meeting point in method.
EDWARD BULWER-LYTTON (1803-1873)

In order to find answers to the research questions, three investigations were carried out, involving administrations of a reading test consisting of both 4- and 3-option items, a survey among content specialists, and a questionnaire completed by the students immediately after having taken the test. The next sections cover the methods used for the collection and analysis of the data for each research question.

3.1 THE EFFECTS ON THE TEST QUALITY

3.1.1 *Instrument*

The first phase of the 3-phase investigation reported in this study explored the possible effects of reducing the number of options on the psychometric properties and completion time of a MC test. The instrument used for collecting the required data was the English Reading Comprehension Test (ERCT), developed to assess proficiency in English reading comprehension of NATO military and civilian personnel who are not native speakers of English. The test consists of 60 MC items covering Levels 1, 2, and 3 in accordance with the proficiency level descriptors of NATO STANAG 6001, with 20 items at each level. Each item contains a reading passage of varying length, followed by a stem (the question) and 4 answer choices. A brief orientation precedes each passage to inform the test taker from what kind of source the passage was selected. All passages, stems, options and orientations are in English, while the test instructions are given in the test takers' native language. Although the actual test questions cannot be provided, illustrative items at the different proficiency levels are given in Appendix 1. The test was developed in 2006 and has been validated in 2007. All MC items for the test were written by trained item writers adhering to high standards of test development and using common item writing principles. All distractors have been written so as to be plausible. The test was designed to be administered online by computer, in which case the items are presented to each candidate in random order within each level. For this study, a paper-and-pencil version of the test has been used.

3.1.2 *Subjects*

The subjects in this study were male and female (18%) military and civilian defence personnel of a large variety of age groups, grades, ranks and branches, and of all services. The participants were randomly selected from a larger group, judged to be representative of the target population of the test. The method of random sampling was used because this strategy is considered more suitable for making generalisations than systematic sampling (Cohen *et al.*, 2007: 110). In order to decrease sampling error, the technique of stratified random sampling (Brown, 2005: 112) was applied, meaning that those characteristics were identified that appear in the wider population which must also appear in the sample, for instance, sex, age, military rank, and educational background. It was assumed that the resulting sample had about the same proportional characteristics as the wider population. However, the selection procedure could not entirely be controlled, because it depended on which students happened to be available at the precise moments that the test was administered. Thus, given that for practical reasons there was some lack of control over the randomisation of exposures, i.e., the access to the sample, the methodology employed for this research project can be termed quasi-experimental (Cohen, 2007: 282).

A possible drawback of many previous studies is that test taker ability was usually not controlled, since the tests in different formats were administered to different groups of examinees (e.g., Green *et al.*, 1982; Trevisan *et al.*, 1991; Crehan *et al.*, 1993; Cizek *et al.*, 1994, 1998; Shizuka *et al.*, 2006). Therefore, a difference in general ability level between examinee groups could have accounted for an unknown proportion of the observed changes in item performance. In other studies (for instance, Landrum *et al.*, 1993; Delgado & Prieto, 1998) a repeated measures design was applied in which two versions of a test were administered to the same group of test takers with a time lag between administrations. Such a design also allows for confounding of results, in this case due to memory of item responses or to learning between test administrations. The present study provided an extension and improvement of the earlier applied methods by giving of a test consisting of both 3-option and 4-option items to the same group of test takers in a single testing occasion. The adopted design constitutes a better setting for the analysis of the research question, since this gives more appropriate indicators to compare the test accuracy according to the ability level being measured, the estimates of the parameters being independent of the sample variations.

3.1.3 *Assembling two versions of the test*

The following procedure for assembling two versions of the test was adopted. Using the item difficulty indices obtained from previous test administrations, all the test items were arranged in order of difficulty per level. The entire test was then divided into two equal halves, the first half containing all odd-numbered items, the second half all even items (10 items per proficiency level). Next, the items of both halves were rearranged in random order within each level. Subsequently, on the basis of previously obtained distractor analysis data, the 4-option items were reduced to a 3-option format by discarding the least frequently chosen distractor. The rationale for using this method of distractor removal will be given below at the discussion of the second research question. In case two distractors had an equal lowest *p*-value, one of these was discarded randomly; this occurred in 7% of the cases (4 items).

In this way, four different test parts were obtained: Part 1, consisting of the 30 odd 4-option items; Part 2 with the 30 even 4-option items; Part 3 with the 30 odd 3-option items; and Part 4 with the 30 even 3-option items. Parts 1 and 3 contained the same items in the same order and differed solely in the number of alternatives per item; the same applied to Parts 2 and 4. Then two test forms were assembled, printed in separate booklets, by combining Parts 1 and 4 in Form A, and Parts 2 and 3 in Form B. Care was taken that in both forms the items were key balanced.

As such, this design guaranteed that in both forms the same test content was sampled, and that all test takers were exposed to both 3- and 4-option items during one testing occasion. In addition, by placing Part 3 before Part 2, Form B started with the 3-option items, whereas Form A began with the 4-option items. This form of counterbalancing ensured that any practice, fatigue or boredom effects on the candidates' performance could not be attributed to the specific item format. Finally, 6 items (10% of all items) consisted of the same number of options in both forms, serving as anchors to compare potential differences in ability between the two groups of test takers. Unlike Straton and Catts (1980), but like Costin (1970, 1972), Ramos and Stern (1973), Trevisan *et al.* (1991, 1994), and others, the number of choice points across forms was not constant. Instead the number of items was kept constant to facilitate the comparison of the completion times of the respective forms, even though it was expected that this procedure would make a direct comparison of the reliability of the test forms less straightforward.

3.1.4 *Method of data collection*

The two forms of the test were administered in multiple sessions to a total of 124 students. No criterion-referenced data were available regarding the English proficiency of these students. At each session, Forms A and B were distributed alternately, no two students sitting next to each other receiving the same form. Normal testing procedures were used except that students were instructed to record on their answer sheets the exact time that they started and the time that they completed each of the two parts of the test. The participants were told that this was a mock exam, but that it would give them a very good estimate of their abilities. It was anticipated that by presenting this test as a "trial" examination a realistic setting would be provided where motivation to do well would be high while avoiding ethical difficulties. The test takers were asked to read each passage and then answer the corresponding MC item using information contained within the passage. They were instructed to answer all items, and they were informed that there would be no penalty for incorrect responses. Although the students were told to work at the same speed as they would normally do taking a reading test, no time limit was imposed to ensure that all test takers answered all the items, thus preventing that the test results and completion times were being influenced by lack of time at the end of the test.

3.2 JUDGING DISTRACTOR EFFECTIVENESS

The aim of the second phase of this 3-phase investigation was to establish whether content specialists are able to judgementally (i.e., without statistical data) identify non-functioning distractors in 4-option items, and to what extent they exhibit mutual agreement in their judgements.

3.2.1 *Method of distractor removal*

In order to clarify what is meant by "non-functioning" in this context, it is necessary to have a closer look at the different methods that can be applied to remove a distractor. In some of the earlier studies investigating the effects of the number of options, the distractor was discarded randomly (e.g., Costin, 1970; 1972; 1976; Straton & Catts, 1980). An obvious downside of this method is that there is a considerable chance that a potentially highly effective distractor will be deleted, which would unintentionally render the item weaker in terms of discriminative power. Lord (1977) argued that dropping the least discriminating option should yield better results than simply eliminating distractors at random. This practice is followed in more recent studies (*cf.*, for instance, Trevisan *et al.*, 1991; Crehan *et al.*, 1993). In most other studies using a removal method on an empirical basis, the least frequently chosen distractor is deleted (e.g., Cizek & O'Day, 1994; Rogers & Harley, 1999; Shizuka *et al.*, 2006).

In the present study, the method of deleting the least frequently chosen distractor was used to reduce the 4-option items to 3-option items of the MC reading test. The reason for this is that this procedure arguably resembles most the method applied by judges to determine the non-functioning option on a non-empirical basis, presuming that it is more difficult to intuitively detect the least discriminating than the least plausible or least attractive distractor. Therefore, the method of deleting the least frequently chosen distractor would lend itself better for a comparative analysis.

Haladyna and Downing (1993) used the term *functioning* distractor to describe the frequency with which option are chosen by test takers. A good distractor should be selected by low achievers and ignored by the rest of the examinees, which presupposes that it must be selected by at least some students, i.e., that it is minimally plausible: 'If less than 5% of all examinees choose it, the distractor is probably so implausible that it probably appeals only to those making random guesses' (Haladyna & Downing, 1993: 1005). From this point of view, options that are selected by fewer than 5% of all test takers would be called *non-functioning distractors*. However, it should be acknowledged that the term "non-functioning distractor" is not entirely appropriate to indicate the least frequently chosen or least plausible distractor. The fact that a distractor is proportionally least frequently endorsed does not make it automatically non-functioning; in actuality, it would still be effective if it were chosen by a considerable (albeit the smallest) number of low achievers and by none of the high achievers. Hence, for the purpose of this study, the term "least functioning distractor" (rather than "non-functioning distractor") is used to refer to the, empirically or judgementally determined, distractor that attracts the smallest number of test takers, regardless of its discriminative power.

3.2.2 *Method of data collection*

Answers to the second research question were sought by asking a group of 14 judges to independently review the complete set of sixty 4-option items of the ERCT and to indicate per item which distractor would probably be chosen least frequently by the test takers, that is, which distractor in their view was least plausible or attractive. Nine of these judges are qualified MC item writers who have been working on the development of the test under consideration. However, none of these item writers had any statistical information about the items at his or her disposal. The remaining five judges are experienced teachers of English who have not been involved in the development of the test and were not familiar with its content.

The findings of the judges were taken together in order to establish the judgementally identified least attractive distractor per item. The individual answers of the judges were used to examine the inter-rater reliability, i.e., the extent that the subject matter experts mutually agreed on their judgements. Finally, using item difficulty and distractor analysis data from both the administrations of the 4-option test parts for this study and from previous administrations of this test, the empirically defined least functioning distractors were identified and compared with the non-empirical data.

3.3 TEST TAKERS' PERCEPTIONS AND PREFERENCES

3.3.1 *Instrument and sample*

The last phase of this study examined the effects of reducing the number of options on the test takers' perceptions and attitudes toward the different formats, and whether there are any differences between high-level students and low-level students in this respect. Data for this research question would be collected from all the test takers participating in the study. Given the number of subjects (*n*=124), a questionnaire was considered the most appropriate means to gather the data. The advantages of using a questionnaire are that substantial amounts of information can be collected in a relatively short time, that managing the information is usually easier than with interviews, and that a certain degree of anonymity can be provided (Banerjee, 2004: 30). A potentially low response rate is often regarded as one of the major disadvantages of questionnaires, but since the questionnaire was to be completed by the test takers immediately after the test, this factor was not considered to play an important role.

3.3.2 *Questionnaire design*

The questionnaire was specifically designed for this study and piloted on a small group of students in order to check that the questions were clear and interpreted similarly by the respondents. It consisted of three sections. The first section contained a small number of questions relating to the respondents' biographical data and their educational background. The second section consisted of 8 closed questions to determine the possible relationship between test takers' perceptions of the 3-option item format and test performance. Closed questions were judged to be the most appropriate question type here because of their suitability to summarise replies to produce a general picture of the sample.

The questions covered four specific areas of interest – (perceived) relative difficulty, reliability, efficiency and suitability. A multi-item scale (Dörnyei, 2003: 33), using differently worded questions focusing on the same target, was employed to reduce the impact of inconsistent responses to one question. Further, in order to minimize the respondents' possible tendency to answer cursorily, the questions were posed in such a way that they represented a mixture of positive and negative attitudes. For the closed questions a 5-point Likert scale was used, ranging from "strongly disagree" to "strongly agree". The midpoint on the scale was labelled "neither disagree nor agree" and represented neutrality towards either option format. In addition, every question offered the possibility to answer with "I don't know", representing undecidedness about the proposition. The third and last section of the questionnaire consisted essentially of a single open question about the test taker's overall preference of option format. Here, the respondents were asked to give an explanation for their preference, allowing them to decide what (additional) information they would like to provide. The questionnaire items are provided in Appendix 2.

CHAPTER 4: RESULTS

Numbers are like people;
torture them enough and they'll tell you anything.
ANONYMOUS

4.1 THE EFFECTS ON THE TEST QUALITY

4.1.1 *Preliminary analyses*

Prior to examining the effects of reducing the number of options on the psychometric properties of the MC reading test, it was necessary to verify whether the two groups of test takers were of equal language ability, and whether the split-half method applied to create test Forms A and B resulted in genuinely parallel test parts. If the general ability level between test taker groups or the difficulty of the respective test parts differed greatly, this could confound any observed changes in item performance.

As a first step, the descriptive and inferential statistics for both test forms were calculated in order to check if the distribution of test scores was normal. Item difficulty, item discrimination, and test reliability indicators in all analyses were calculated from a Classical Test Theory methodology, given that the sample size per test (*n*=62) would not allow the appropriate parameter estimation if using the Item Response Theory approach. Examination of the descriptive statistics (see Appendix 3), and box and whisker plots revealed that the distributions of the two sets of scores were approximately normally distributed and that there were no extreme scores.[1] It should be noted, though, that the comparatively high *p*-values of both test forms (Form A: *p* =.78; Form B: *p* = .79) indicate that this was a fairly easy test for these particular groups of test takers, and that the sample as such was somewhat truncated.

Next, possible differences in language ability between the two groups of test takers (Group 1, who took Form A, and Group 2, who took Form B) were examined. An independent-samples *t*-test assuming equal variances was conducted to compare the mean scores on the 6 anchor items which were in both forms identical in every respect, including the number of options. There was no significant difference in ability between Group 1 (*M*=4.52, *SD*=1.11), and Group 2 (*M*=4.71, *SD*=1.23, *t*(122)=-.918, *p* =.361, 2-tailed).

[1] All statistical analyses in this study were carried out using SPSS 14.0.

The comparability of both test parts of each form was checked by conducting independent *t*-tests on mean scores for Parts 1 and 2 (each consisting of 4-option items) and for Parts 3 and 4 (each with 30 3-option items). There were no significant differences found in difficulty between the 4-option test parts in Form A (*M*=23.66, *SD*=3.858) and Form B (*M*=23.81, *SD*=3.584, *t*(122)=-.217, *p*=.829, 2-tailed), nor between the 3-option test parts in Form A (*M*=23.60, *SD*=3.523) and Form B (*M*=23.79, *SD*=3.640, *t*(122)=-.301, *p*=.764, 2-tailed).

Finally, the relationship between the various test halves was investigated using Pearson product-moment correlation coefficient. There was a strong, positive correlation between the test parts of Form A (*r* =.75, *n*=62, *p*<.01) and a similar correlation between the test parts of Form B (*r* =.69, *n*=62, *p*<.01).

On the basis of these statistics it is safe to conclude that, on average, the two groups of test takers were of equal ability and that the two test parts on both forms were indeed equivalent and measured the same property. This means that possible changes in item performance might be attributed most confidently to the change in option format.

4.1.2 *Data triangulation*

The actual comparison of the 4-option test items and the 3-option items was done by data triangulation. The purpose of data triangulation is to enhance the concurrent validity of the measurement findings by using multiple comparative analyses (Cohen, 2007: 144). The design of the present study permitted both a *between-tests* and a *within-tests* examination of the effects of the number of options on the item performance. In the between-tests examinations the performance of *identical items* with different numbers of options and different student samples were compared, whereas in the within-tests method items with different numbers of options and *identical student samples* were examined.

In order to determine whether reducing the number of options affected students of differing ability differently, the test takers of both Group 1 and Group 2 were grouped into high-, middle- and low-ability levels according to their performance on the test. The group of high achievers were the 27% of test takers having obtained highest total test scores, and the group of low achievers consisted of the 27% of test takers with the lowest scores. The upper and lower 27% were chosen because this percentage provides the best compromise between two desirable but incompatible aims in comparing student performance: (1) to make the extreme groups as large as possible, and (2) to make the extreme groups as different as possible (Ebel, 1972: 385). Data were available for 124 test takers, so the upper and lower groups consisted each of 34 test takers (17 each in Group 1 and Group 2).

Criteria on which the 3-option and 4-option format contrasts were compared include: mean test score and item difficulty, mean item discrimination (point biserial correlation), estimates of reliability (Cronbach's alpha) and the associated standard error of measurement. In addition, the mean test completion times and the mean times per item were calculated. For reasons explained in Appendix 4, it was decided not to correct the scores for guessing. Item analyses were run on each part of each test form in order to obtain the required data. A summary of the observed psychometric characteristics of Parts 1 and 4 (Form A), and Parts 2 and 3 (Form B) is presented in Table 4.1.

Table 4.1: Characteristics separate test parts

		FORM A		**FORM B**	
		Part 1	**Part 4**	**Part 2**	**Part 3**
Statistic		4-option items	3-option items	4-option items	3-option items
Sample size		62	62	62	62
No. of items (= max. score)		30	30	30	30
Mean score	*Total*	23.66	23.58	23.81	23.79
	*Lower group **	18.88	19.35	18.94	20.41
	*Upper group **	27.71	27.47	26.94	27.88
Std. deviation		3.858	3.504	3.540	3.614
Mean item difficulty		.79	.79	.79	.79
Mean discrimination (r_{pbi})		.24	.21	.20	.26
Reliability (α)		.75	.67	.69	.71
SEM		1.93	2.01	1.97	1.95

** n=17*

4.1.3 *Mean test scores and item difficulty indices*

From Table 4.1 it can be seen that the mean scores, and consequently the mean difficulty indices, of the respective test parts were almost identical for the total groups. Paired and independent *t*-tests were run on the mean scores for the respective test parts to examine whether the differences were statistically significant. The results of these tests, summarized below in Table 4.2, show that all differences were non-significant and that the null hypothesis was supported in all cases.

Table 4.2: Summary of results of t*-tests on mean scores separate test parts*

	ALL ABILITY GROUPS	**3-option items**	
		Part 3 *(items 1-30)* *M* = 23.79 *SD* = 3.640	**Part 4** *(items 31-60)* *M* = 23.60 *SD* = 3.523
4-option items	**Part 1** *(items 1-30)* *M* = 23.66 *SD* = 3.858	*t* (122) = -.192 (n.s.) *p* = .848 (2-tailed)	*t* (61)= .194 (n.s.) *p* = .847 (2-tailed)
	Part 2 *(items 31-60)* *M* = 23.81 *SD* = 3.584	*t* (61)= .044 (n.s.) *p* = .965 (2-tailed)	*t* (122)= -.329 (n.s.) *p* = .743 (2-tailed)

When run for the groups of low- and high-ability students separately, the *t*-tests yielded the following results (see Tables 4.3 and 4.4):

Table 4.3: Summary of results of t*-tests on mean scores low-ability groups*

	LOW-ABILITY GROUPS	**3-option items**	
		Part 3 *(items 1-30)* *M* = 20.41 *SD* = 3.554	**Part 4** *(items 31-60)* *M* = 19.35 *SD* = 2.060
4-option items	**Part 1** *(items 1-30)* *M* = 18.88 *SD* = 2.848	*t* (32) = -1.385 (n.s.) *p* = .176 (2-tailed)	*t* (16)= -.548 (n.s.) *p* = .591 (2-tailed)
	Part 2 *(items 31-60)* *M* = 18.94 *SD* = 2.331	*t* (16)= -1.966 (n.s.) *p* = .067 (2-tailed)	*t* (32)= .546 (n.s.) *p* = .589 (2-tailed)

Table 4.4: Summary of results of t*-tests on mean scores high-ability groups*

HIGH-ABILITY GROUPS		**3-option items**	
		Part 3 *(items 1-30)* *M* = 27.88 *SD* = 1.364	**Part 4** *(items 31-60)* *M* = 27.47 *SD* = 1.125
4-option items	**Part 1** *(items 1-30)* *M* = 27.71 *SD* = 1.359	*t* (32) = -.378 (n.s.) *p* = .708 (2-tailed)	*t* (16)= .578 (n.s.) *p* = .571 (2-tailed)
	Part 2 *(items 31-60)* *M* = 26.94 *SD* = 1.853	*t* (16)= -1.628 (n.s.) *p* = .123 (2-tailed)	*t* (32)= 1.007 (n.s.) *p* = .321 (2-tailed)

Although the groups of both low and high achievers scored slightly better on the 3-option test parts, none of the differences between any of the test parts were statistically significant. From this it can be deduced that the option format had virtually no meaningful effect on the performance of the test takers, regardless of their ability level.

4.1.4 *Reliability*

The index of reliability used here is Cronbach's alpha coefficient. Alpha measures the extent to which item responses obtained at the same time correlate highly with each other. The standard error of measurement (SEM) is closely related to coefficient alpha and can be interpreted as a standard deviation of the test taker's score across multiple administrations (Bachman, 1990: 199). The smaller the SEM, the higher the precision of interpreting any particular score as being representative of the test taker's true score.

The observed reliability coefficients for the test parts with identical items (Parts 1-3, and Parts 2-4) were somewhat lower in the 3-option format (.04 and .02, respectively). Table 4.1 also shows that the SEM increased for the 3-option test parts, indicating that the scores are slightly less accurate. Although these differences are so marginal that they might be the result of sampling error, they could also be explained by the fact that test reliability depends to some extent on score variability (Ebel, 1972: 430). Score variability, in turn, is related to the effective range of scores. The effective range is the maximum possible range minus the expected chance score. A 4-option test has a chance score of 25%, whereas the chance score of a 3-option test is 33%. Thus, each of the 4-option parts of the ERCT here has an effective range of 22.5 (30–(.25 x 30)), whereas each 3-option test part has an effective range of 20 (30–(.33 x 30)). Theoretically, the reduced score variability as a result of the smaller effective range would lead to a lower reliability coefficient for a 3-option test. However, as can be seen from Table 4.1, the standard deviations for all test parts are approximately the same, or at least not consistently lower. This suggests that the reduction of the number of options had a much smaller effect on the score variability than expected. Studies by Haladyna and Downing (1993), and Shizuka *et al.* (2006), discussed in Chapter 2, showed that a possible explanation for the fact that test reliability is hardly affected by a reduction of the number of options may be found in an analysis of the distractors.

4.1.5 *Item discrimination*

Before looking in more detail at the distractor effectiveness, it should be pointed out that discrimination is partly a function of the differences in ability levels among the test takers (Henning, 1987: 53). The more homogeneous the ability of the test takers (i.e., the more narrow the ability range), the less test items will discriminate between high and low achievers. Further, as explained in Ebel (1972: 390), items with a *p*-value of .50 yield potentially maximum discrimination power. Very hard or very easy items (with *p*-values below .20 and above .80) supply less than half of the maximum potential differential information. In the present study more than 50% of the items had *p*-values above .80. In other words, the relatively high average ability level of the test takers had a deteriorating effect on the calculated point biserial correlations.

However, of interest here are not so much the values for the observed point biserial coefficients in absolute terms, as the differences in these values relative to the item format. Table 4.1 shows that in both cases the test parts with 3-option items had somewhat higher discriminations than their 4-option counterparts. Contrary to expectations, the 3-option format discriminated slightly better between upper and lower ability students despite having one option less. Independent *t*-tests were run on the point biserial correlation (r_{pbi}) coefficients of test parts containing the same items but with different numbers of options (i.e., Parts 1-3, and Parts 2-4). There were no significant differences found in the mean item discrimination between Part 1 (*M*=.243, *SD*=.180) and Part 3 (*M*=.265, *SD*=.165; *t*(58)=-.470, *p*=.640, 2-tailed), or between Part 2 (*M*=.197, *SD*=.134) and Part 4 (*M*=.207, *SD*=.168; *t*(58)=-.258, *p*=.798, 2-tailed).

In order to find an answer to the question why item discrimination between 3-option and 4-option items did not differ significantly, the r_{pbi} coefficients of the individual items were subjected to a closer examination. For the purpose of evaluation the point biserial coefficients were categorized as follows: A poor discrimination index was defined as a point biserial that falls in the range between less than zero to .09; fair, any value between .10 to .29; and good, values equal to or greater than .30 (Hopkins, 1998: 260). Further, it is important to recall that items 1-10 and 31-40 were relatively easiest (at Level 1 according to NATO STANAG 6001), and items 21-30 and 51-60 comparatively hardest (at Level 3). Inspection of the data in Table 4.5 on page 32 revealed that although the differences in the mean item discrimination indices at the test parts level were very small, the differences at the item level were sometimes considerable. In 43 out of 60 items (72%) the discrimination index of the 3-option items was more than .05 lower or higher than the discrimination index of the 4-option items, and in 10 cases (17%) the difference was even more than .30.

In addition, several trends among the item r_{pbi} coefficients were detected. The first trend was that the number of items classified as good discriminators increased as the number of options decreased. More precisely, the 3-option test parts had 7-10% more items that highly discriminated than the 4-option test parts. The second trend was that, overall, the number of good discriminators increased as the difficulty level of the items increased. A final trend was that, on average, the r_{pbi} coefficients decreased as the difficulty level of the items increased and the number of options decreased.

Table 4.5: Point biserial correlation coefficients of separate test parts

Item No.	Part 1 *(4-options)*	Part 3 *(3-options)*	Difference	Item No.	Part 2 *(4-options)*	Part 4 *(3-options)*	Difference
1	.141	–.098	*–.239*	*31*	.000	.341	*.341*
2	.000	.261	*.261*	*32*	.000	–.161	*–.161*
3	.300	.408	*.108*	*33*	.124	.283	*.159*
4	.011	.433	*.422*	*34*	.002	–.236	*–.238*
5	.082	.274	*.192*	*35*	.334	.327	*–.007*
6	.405	.436	*.031*	*36*	.021	–.027	*–.048*
7	.032	.052	*.020*	*37*	.179	.170	*–.009*
8	–.078	.324	*.402*	*38*	–.016	.000	*.016*
9	.190	.538	*.348*	*39*	.157	.170	*.013*
10	–.006	.280	*.286*	*40*	.139	.280	*.141*
11	.094	.317	*.223*	*41*	.123	.117	*–.006*
12	.157	.383	*.226*	*42*	.327	.375	*.048*
13	.184	.272	*.088*	*43*	.082	.283	*.201*
14	.206	.423	*.217*	*44*	.264	.096	*–.168*
15	.498	.331	*–.167*	*45*	.087	.327	*.240*
16	.407	.420	*.013*	*46*	.198	.292	*.094*
17	.211	.217	*.006*	*47*	.319	.345	*.026*
18	.138	.232	*.094*	*48*	.343	.204	*–.139*
19	.400	.182	*–.218*	*49*	.283	.393	*.110*
20	.148	.044	*–.104*	*50*	.202	.223	*.021*
21	.511	.492	*–.019*	*51*	.440	.336	*–.104*
22	.465	.168	*–.297*	*52*	.153	.092	*–.061*
23	.241	.330	*.089*	*53*	.245	.116	*–.129*
24	.534	.189	*–.345*	*54*	.208	.372	*.164*
25	.093	.325	*.232*	*55*	.434	.058	*–.376*
26	.535	.193	*–.342*	*56*	.425	.112	*–.313*
27	.334	.331	*–.003*	*57*	.240	.174	*–.066*
28	.464	.158	*–.306*	*58*	.079	.340	*.261*
29	.330	–.213	*–.543*	*59*	.330	.334	*.004*
30	.291	.245	*–.046*	*60*	.198	.487	*.289*
M	**.244**	**.265**	***.021***	*M*	**.197**	**.207**	***.010***
Poor	**8 (27%)**	**4 (13%)**	***–14%***	*Poor*	**8 (27%)**	**7 (23%)**	***–4%***
Fair	**10 (33%)**	**12 (40%)**	***7%***	*Fair*	**14 (46%)**	**12 (40%)**	***–6%***
Good	**12 (40%)**	**14 (47%)**	***7%***	*Good*	**8 (27%)**	**11 (37%)**	***10%***

These findings can be verified by categorizing the items that had one or more distractors chosen by at least 5% of the test takers, – often used as the minimum endorsement frequency to consider a distractor functional (e.g., Haladyna & Downing, 1993: 1005; Cizek & O'Day, 1994: 865). Table 4.6 presents the number of 4- and 3-option items with minimally discriminating distractors at the different STANAG levels: Level 1 (*easy*), Level 2 (*average*) and Level 3 (*hard*).

Table 4.6: Number of items with discriminating distractors

All ability groups	**4-option items** *Parts 1 and 2*				**3-option items** *Parts 3 and 4*			
No. of items with…	*Easy*	*Avg.*	*Hard*	*Total*	*Easy*	*Avg.*	*Hard*	***Total***
***3** discriminating distractors*	0	2	8	**10**	n/a	n/a	n/a	n/a
***2** discriminating distractors*	1	6	12	**19**	1	7	18	**26**
***1** discriminating distractor*	4	7	0	**11**	6	9	2	**17**
***0** discriminating distractors*	15	5	0	**20**	13	4	0	**17**
Total items	20	20	20	60	20	20	20	60
Total number of distractors				180				120
Total number (percentage) of discriminating distractors				**79** *(43.9%)*				**69** *(57.5%)*
Mean number of discriminating distractors per item				**1.32**				**1.15**

(n=124)

Table 4.6 shows that the Level 3 items had considerably more functional distractors than the Level 1 items, but also that the total percentage of functional distractors increased by more than 13% as the number of options decreased. Another observation was that items with 3 effectively functioning distractors were quite rare, especially at the lower difficulty levels. The distractor analysis revealed further that irrespective of the number of options per item, the mean number of functioning distractors was much lower than 2.

However, the numbers of items with discriminating distractors by and of themselves do not reveal much about qualities of the options. Of importance is not only how many but also what level of the test takers chose which options. After the total group of test takers had been split by ability level, the picture became somewhat more differentiated (see Tables 4.7 and 4.8).

Table 4.7: Number of items with discriminating distractors high-ability groups

High-ability groups *(n=34)*	**4-option items** *Parts 1 and 2*				**3-option items** *Parts 3 and 4*			
No. of items with...	*Easy*	*Avg.*	*Hard*	*Total*	*Easy*	*Avg.*	*Hard*	***Total***
***3** discriminating distractors*	0	0	1	**1**	n/a	n/a	n/a	n/a
***2** discriminating distractors*	0	3	10	**13**	1	2	6	**9**
***1** discriminating distractor*	6	6	6	**18**	4	10	9	**23**
***0** discriminating distractors*	14	11	3	**28**	15	8	5	**28**
Total items	20	20	20	60	20	20	20	60
Total number of distractors				180				120
Total number (percentage) of discriminating distractors				**47** *(26.1%)*				**41** *(34.2%)*
Mean number of discriminating distractors per item				**.78**				**.68**

Table 4.8: Number of items with discriminating distractors low-ability groups

Low-ability groups *(n=34)*	**4-option items** *Parts 1 and 2*				**3-option items** *Parts 3 and 4*			
No. of items with...	*Easy*	*Avg.*	*Hard*	*Total*	*Easy*	*Avg.*	*Hard*	***Total***
***3** discriminating distractors*	0	8	13	**21**	n/a	n/a	n/a	n/a
***2** discriminating distractors*	6	10	7	**23**	4	16	18	**38**
***1** discriminating distractor*	12	2	0	**14**	13	4	2	**19**
***0** discriminating distractors*	2	0	0	**2**	3	0	0	**3**
Total items	20	20	20	60	20	20	20	60
Total number of distractors				180				120
Total number (percentage) of discriminating distractors				**123** *(68.3%)*				**95** *(79.2%)*
Mean number of discriminating distractors per item				**2.05**				**1.58**

From these Tables it can be observed that language ability was adequately reflected in the option selection. High language ability should correlate highly with choosing the correct option (hence the great number of items without discriminating distractors in Table 4.7), whereas low language ability is expected to correlate highly with choosing any of the distractors. For the high achievers, there was only a moderate increase in the mean number of discriminating distractors per item (from .68 to .78) as the number of options increases. In fact, of the 60 additional distractors in the 4-option test parts, 54 (90%) were not chosen at all, or at best chosen by random guessers. In other words, the fourth option did not contribute much to the discriminatory power of the test.

For the low achievers, the mean number of discriminating distractors per item was considerably higher in the 4-option test parts than in the 3-option parts. The percentage of items with 3 effectively performing distractors was more than 30%. Almost half of the 60 additional distractors in the 4-option test parts were chosen by the low achievers. When presented with 4 options, test-takers' actual choices spread out over a much larger range – over more than 3 options – than when given 3 options per item.

On the basis of these results, it is apparent that the earlier observed non-significant differences in mean discrimination between the 4-option and 3-option test parts is not a result of a systematic lack of differences in the point biserial coefficients at the item level. On the contrary, at times reducing the number of options led to considerable changes in the coefficients at the varying difficulty levels. However, these changes tended to go in opposite directions (generally increasing at the lower difficulty levels and decreasing for harder items), and had thus a compensatory effect on the total mean discrimination indices.

4.1.6 *Completion time*

Table 4.9 displays the mean completion times per test part. As this Table indicates, mean time taken to complete the respective parts increased with the number of options per item. This was true regardless of student ability level. It seems then that time to completion is positively related to the number of options per item.

A statistical analysis was performed to check whether the differences in mean completion times of the test parts were significant. Given that the mean item difficulty values and the mean length of the reading passages were almost identical for all test parts, the four test parts were considered sufficiently equivalent to conduct a one-way between groups analysis of variance to explore the impact of the number of options on the test completion time. Analysis of variance found that there was a statistically significant difference in completion times between the four test parts (*F*(3, 244)=9.136, *p*<.001). The post-hoc comparisons using the Tukey HSD test indicated that the means for Part 1 (*M*=61.45, *SD*=12.942), Part 2 (*M*=62.71, *SD*=12.574) and Part 3 (*M*=60.79, *SD*=11.914) were not statistically significantly different from each other. However, the Tukey test found that the mean completion time for Part 4 (*M*=52.50, *SD*=10.586) was distinctly and significantly different at the *p*<.05 level from all other three test parts.

Table 4.9: Mean completion times

		FORM A		FORM B	
		Part 1 4-option items	**Part 4** 3-option items	**Part 2** 4-option items	**Part 3** 3-option items
Sample size		62	62	62	62
No. of items		30	30	30	30
Mean length of reading passages *(words)*		142	148	148	142
Mean completion time *(min.)*	*Total*	61.45	52.50	62.71	60.79
	*Lower group **	66.76	56.82	65.88	65.47
	*Upper group **	58.18	50.06	59.12	55.00
Mean time per item *(min.)*		2.05	1.75	2.09	2.03

** n=17*

The results with regard to completion time are thus mixed. On the one hand there is a small, non-significant difference in mean completion time between Parts 1 and 3 (containing identical items except for number of options), which would result in negligible savings in administration time using 3-option items. On the other hand, the difference in mean completion time between Parts 2 and 4 (each also with the same items) is not only statistically significant, but also substantial in terms of gains in testing time. There is no obvious explanation for the fact that the completion time of Part 4 is so much shorter than the other 3-option test part, especially considering that on average the reading texts are even longer in Part 4. Most probably it has to be attributed to boredom or practice effects, which makes that test takers read and answer questions faster towards the end of the test. This would also explain the relatively small difference in completion time between Part 3 and Part 2 (which came last in Form B): although the 3-option test part is completed faster than the 4-option items, boredom and/or practice effects may have attenuated these differences.

In order to get a more clear-cut picture of the effect of the number of options on the completion time, the mean times of the 4-option test parts were averaged and compared to the averaged mean completion times of the 3-option parts. The average completion time of the 4-option items was 62.08 minutes, and that of the 3-option items 56.65 minutes. This difference implies that, using the longer time of the 4-option items as a standard, about 9% more 3-option items could be squeezed in. Those extra items would enhance both content validity and reliability.

4.1.7 *Summary of results*

The number of options had no statistically significant effect on the performance of the test takers, irrespective of their ability level: low achievers did comparatively no better or worse on the 3-option items than high achievers. The reliability estimates were marginally lower for the 3-option test parts, which might be caused by the slightly reduced score variability associated with the 3-option format. Despite having one option less, the 3-option test items discriminated, on average, somewhat better between high- and low-ability students. Distractor analysis revealed that this might be explained by the fact that few 4-option items had 3 effectively functioning distractors and that in most cases the fourth option did not contribute to the discrimination of the item. Finally, a one-way ANOVA revealed that the completion time of one test part consisting of 3-option items was significantly (p<.05) shorter than the other test parts. The average completion time of the 3-option test parts was approximately 9% shorter than the 4-option test parts.

4.2 JUDGING DISTRACTOR EFFECTIVENESS

4.2.1 *Degree of agreement*

The aim of the second research question was to establish to what extent the judgementally (i.e., without statistical data) identified least frequently chosen distractors in 4-option items match with those based on the actual statistical performance of the items. A total of fourteen judges participated in this part of the study, nine of whom were trained item writers familiar with the contents of the ERCT but not with the actual student sample. In the analysis, these judges will be referred to as the "content experts". Of the remaining five content specialist, three were familiar with the student sample in this study but not with the actual reading test (hereafter the "sample experts"), and two judges were teachers of English not familiar with either the test or the sample (henceforth called the "lay judges").

First, the least frequently chosen distractors were empirically identified using distractor analysis data from previous test administrations (n=102) and from the actual administrations of the ERCT for the purpose of this study (n=62).[2] Next, the empirical data were compared with the intuitive judgements of the experts using Cohen's kappa. The kappa statistic measures the degree of agreement between the variables above that

[2] The data collected at previous administrations were from the online version, which could have resulted in slight differences in item performance due to modes of delivery.

expected by chance alone; as such it is a more robust and conservative measure than simple percent agreement calculation. It has a maximum of 1 when agreement is perfect, 0 when agreement is no better than chance, and negative values when agreement is worse than chance. Although there are no absolute cut-offs for kappa coefficients, Landis & Koch (1977: 165) suggest the following guidelines: .00–.20 slight agreement; .21–.40 fair; .41–.60 moderate; .61–.80 substantial; and above .80 almost perfect agreement.

Items with difficulty exceeding .90 were eliminated from further analysis because distractors in these items will seldom or only randomly be selected, rendering distractor analysis meaningless. In total 17 items were discarded following this criterion. Table 4.10 shows an information matrix listing the degree of agreement (*kappa*) between the empirical data and the judges as a group, individually and according to qualification as content expert, sample expert or lay judge.

Table 4.10: Degree of agreement (κ) between judges and empirical data

	Judge ID (*no. of items*=43)	**Empirical data** (*no. of items*=43)	**Group rating** (*no. of items*=37)
		κ	*κ*
Group Combined		.567**	–
Content experts	1	.523**	.555**
	2	.345**	.629**
	3	.197	.346**
	4	.344**	.634**
	5	.444**	.528**
	8	.408**	.555**
	9	.466**	.589**
	11	.280*	.488**
	13	.348**	.524**
	M	*.373*	*.539*
	SD	*.100*	*.087*
Sample experts	6	.503**	.573**
	10	.410**	.561**
	12	.408**	.567**
	M	*.440*	*.567*
	SD	*.054*	*.006*
Lay judges	7	.308*	.524**
	14	.150	.295*
	M	*.229*	*.410*
	SD	*.112*	*.162*
Total	*M*	*.367*	*.526*
	SD	*.108*	*.096*

** Significant at *p*<.001 (2-tailed); * Significant at *p*<.05 (2-tailed)

From the data presented in this Table the following observations can be made:

- The kappa coefficient of the combined group (representing the choice of the majority of the judges) was .57, suggesting a moderate agreement between the judgements of the experts and the empirically determined least functional distractor. The combined group coefficient is higher than the average of the individual judges (κ=.37), but also higher than the coefficient of any of the judges individually. This indicates that using a majority choice resulted in a better prediction about the least attractive distractors than using the judgements of experts individually.
- The average agreement between each individual judge and the majority choice was .53, which suggests that the individual expert judges do agree about the least functional distractor more with each other than with the actual test takers in this study.
- The lay judges showed least agreement with both the test takers and with the other judges.
- On average, sample experts (mean κ=.44) were somewhat better than content experts (mean κ=.37), and much better than lay judges (mean κ=.23) able to predict which distractors are least attractive to test takers. It seems, then, that knowledge of the actual test population increases the reliability of the predictions about the attractiveness of distractors.

4.2.2 *The accuracy of the judgements*

In order to find out which other factors may have played a role in intuitively identifying the least frequently chosen distractor, the exact matches between the choice of the expert judges and the empirically based choices were more closely examined. Table 4.11 displays the item difficulty for each of the remaining 43 items, the empirically observed least frequently chosen distractors (with the percentage choosing those options), the distractors judged as least attractive by the majority of the content experts, and the number of matches. Choices by the judges that did not correspond with the empirical data but with an endorsement percentage of less than 5% were also considered as a match. Where the combined ratings of the judges did not result in a single least attractive distractor, the data were treated as missing values in the analysis.[3]

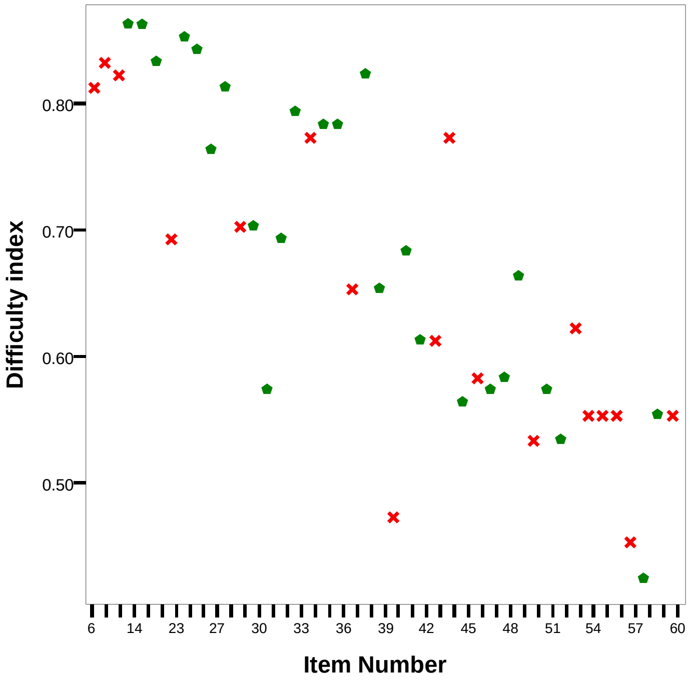
[3] For this part of the study, the test items were arranged by difficulty level, items 1-20 being at Level 1, items 21-40 at Level 2, and items 41-60 at Level 3. For this reason, the item numbers in Table 4.11 do not coincide with those in Table 4.5.

Table 4.11: Least frequently chosen distractors

Level	Item		Least frequently chosen distractor *(percentage choosing)*				Match	Level	Item		Least frequently chosen distractor *(percentage choosing)*				Match
	No.	*p*	*Empirically determined (n=164)*		*Judgementally determined (n=14)*				No.	*p*	*Empirically determined (n=164)*		*Judgementally determined (n=14)*		
1	*6*	*.81*	C	*(4)*	D	*(6)*		**3**	*41*	*.68*	B	*(7)*	B	*(7)*	✓
	9	*.83*	B	*(0)*	C	*(6)*			*42*	*.61*	C	*(9)*	C	*(9)*	✓
	12	*.82*	C	*(3)*	B	*(10)*			*43*	*.61*	A	*(8)*	C	*(15)*	
	14	*.86*	A	*(4)*	A	*(4)*	✓		*44*	*.77*	C	*(5)*	A/C	*(5/11)*	
	15	*.86*	C	*(1)*	C	*(1)*	✓		*45*	*.56*	A	*(12)*	A	*(12)*	✓
	16	*.83*	A	*(1)*	A	*(1)*	✓		*46*	*.58*	B	*(7)*	D	*(19)*	
2	*23*	*.69*	A	*(1)*	D	*(12)*			*47*	*.57*	A	*(1)*	A	*(1)*	✓
	25	*.85*	A	*(5)*	A	*(5)*	✓		*48*	*.58*	D	*(11)*	D	*(11)*	✓
	26	*.84*	C	*(3)*	C	*(3)*	✓		*49*	*.66*	C	*(8)*	C	*(8)*	✓
	27	*.76*	A	*(4)*	A/B	*(4/13)*			*50*	*.53*	B	*(11)*	C	*(15)*	
	28	*.81*	D	*(5)*	D	*(5)*	✓		*51*	*.57*	A	*(7)*	A	*(7)*	✓
	29	*.70*	D	*(5)*	D	*(5)*	✓		*52*	*.53*	D	*(3)*	D	*(3)*	✓
	30	*.70*	B	*(9)*	B	*(9)*	✓		*53*	*.62*	D	*(8)*	C	*(12)*	
	31	*.57*	A	*(6)*	A	*(6)*	✓		*54*	*.55*	D	*(10)*	C/D	*(10/19)*	
	32	*.69*	B	*(1)*	A	*(10)*			*55*	*.55*	B	*(12)*	D	*(16)*	
	33	*.79*	D	*(4)*	D	*(4)*	✓		*56*	*.55*	A	*(11)*	A/C	*(11/18)*	
	34	*.77*	D	*(1)*	D	*(1)*	✓		*57*	*.45*	B	*(7)*	A	*(13)*	
	35	*.78*	C	*(6)*	C	*(6)*	✓		*58*	*.42*	A	*(7)*	A	*(7)*	✓
	36	*.78*	C	*(6)*	C	*(6)*	✓		*59*	*.55*	C	*(14)*	C	*(14)*	✓
	37	*.65*	D	*(4)*	B/D	*(4/17)*			*60*	*.55*	C	*(9)*	D	*(17)*	
	38	*.82*	B	*(5)*	B	*(5)*	✓								
	39	*.65*	A	*(6)*	A	*(6)*	✓								
	40	*.47*	C	*(12)*	B/C	*(13/12)*									

From this Table it appears that the judges do a better job in predicting the least frequently chosen distractors for the Level 2 items (70% matches) than for the Level 3 items (50% matches). One straightforward conclusion would be that the higher the level, the more difficult it becomes to intuitively detect the least frequently distractor. From this argument it follows that the number of matches would be greatest for the Level 1 items. However, this can not be verified with confidence due to the probability of sampling error associated with the small number of Level 1 items retained for this analysis. At the same time, from the item difficulty parameters it can be observed that there does not seem to be any systematic relationship between the difficulty of an item and the probability of detecting intuitively the least attractive distractor. This finding is confirmed by the scatter plot of matches in relation to the *p*-value (see Figure 4.1).

Figure 4.1: Scatter plot of matches (⬟) and non-matches (✖) against item difficulty



Rather than the item's difficulty, the item's quality is likely to be a factor here: surely, a least frequently chosen distractor is for whatever reason less attractive or plausible to the test takers, indicating that it must be of lesser quality than the other distractors. Apparently, the distractor's quality and the probability of being intuitively identified as the least frequently chosen option are inversely related. Put differently, the more flawed the distractor is *in comparison with* the other distractors, the more reliably judges will be able to predict that it attracts fewest test takers. If none or more than one distractor is flawed, the accuracy of the judgements decreases. Consequently, it can be assumed that matches are most likely to occur for items that have only *one* flawed or less plausible distractor. The data in Table 4.12 seem to support this assumption. This Table shows the matches for items with only one distinct least frequently chosen distractor, defined here as a distractor with an endorsement percentage that is at least 5% less than that of the other distractors in the item. The data revealed that in 12 of the 18 non-matches (67%), the item contained none or more than one distinct least frequently chosen distractor, whereas matches occurred for 19 of 25 items (76%) with only one distinct least chosen distractor.

Table 4.12: Matches occurring for items with one distinct least frequently chosen distractor

Level	Item No.	*Match*	ONE distinct least chosen distractor	Level	Item No.	*Match*	ONE distinct least chosen distractor
1	*6*	✗	✗	**3**	*41*	✓	✓
	9	✗	✗		*42*	✓	✓
	12	✗	✗		*43*	✗	✓
	14	✓	✓		*44*	✗	✗
	15	✓	✗		*45*	✓	✓
	16	✓	✓		*46*	✗	✗
2	*23*	✗	✓		*47*	✓	✓
	25	✓	✗		*48*	✓	✓
	26	✓	✓		*49*	✓	✗
	27	✗	✓		*50*	✗	✗
	28	✓	✓		*51*	✓	✓
	29	✓	✓		*52*	✓	✓
	30	✓	✗		*53*	✗	✗
	31	✓	✓		*54*	✗	✓
	32	✗	✗		*55*	✗	✗
	33	✓	✓		*56*	✗	✗
	34	✓	✓		*57*	✗	✓
	35	✓	✗		*58*	✓	✓
	36	✓	✓		*59*	✓	✓
	37	✗	✓		*60*	✗	✗
	38	✓	✗				
	39	✓	✓				
	40	✗	✗				

In light of these findings, it cannot be confidently determined if the observed value of κ=.57 indicates whether (a) judges exhibit a fairly good ability to reliably predict which distractor will be least frequently chosen, or (b) the test contained a relatively large proportion of items with one distinct less frequently chosen distractor. At any rate, it should be emphasized that, even if the latter were the case, this does not mean that the distractors as such are necessarily flawed: every item, no matter how well its distractors are designed, will have a distractor that is chosen by a smallest number of test takers.

4.2.3 *Summary of results*

Statistical analysis of the experts' judgements about the least frequently chosen distractor showed a moderate agreement with the empirical data. The probability of making successful predictions appeared to increase if an item had only one distinct least frequently endorsed distractor. The analysis showed further that using a majority choice resulted in a better prediction about the least attractive distractors than using the judgements of experts individually, and that background knowledge of the actual test population enhances the trustworthiness of the judgements. However, any conclusions based on a sample size as small as the one used in this investigation are vulnerable to error and must therefore be considered tentative without further support.

4.3 TEST TAKERS' PERCEPTIONS AND PREFERENCES

The third and last phase of this study involved an investigation of the test takers' attitudes and preferences with regard to the 3-option format, and whether there are any differences between high-level students and low-level students in this respect. For purposes of analysis, the original 8 closed questions in the questionnaire (see Appendix 2) were clustered into 4 categories. Each category focussed on one of the following sub areas: difficulty, reliability, efficiency and suitability of the 3-option format as perceived by the test takers. As 7 test takers failed to complete the questionnaire, data of 117 respondents were used in this part of the study. Due to the comparatively small sample size, chi square tests could not be performed because the assumption of the minimum expected cell frequency (≥ 5) was violated. Therefore only observed, and not expected counts are reported. Figures 4.2 through 4.5 present pie charts of the answers in each category for the entire group of test takers, and for the low (*n*=33) and high (*n*=32) achievers separately. Omitted answers were placed under the "no opinion" position.

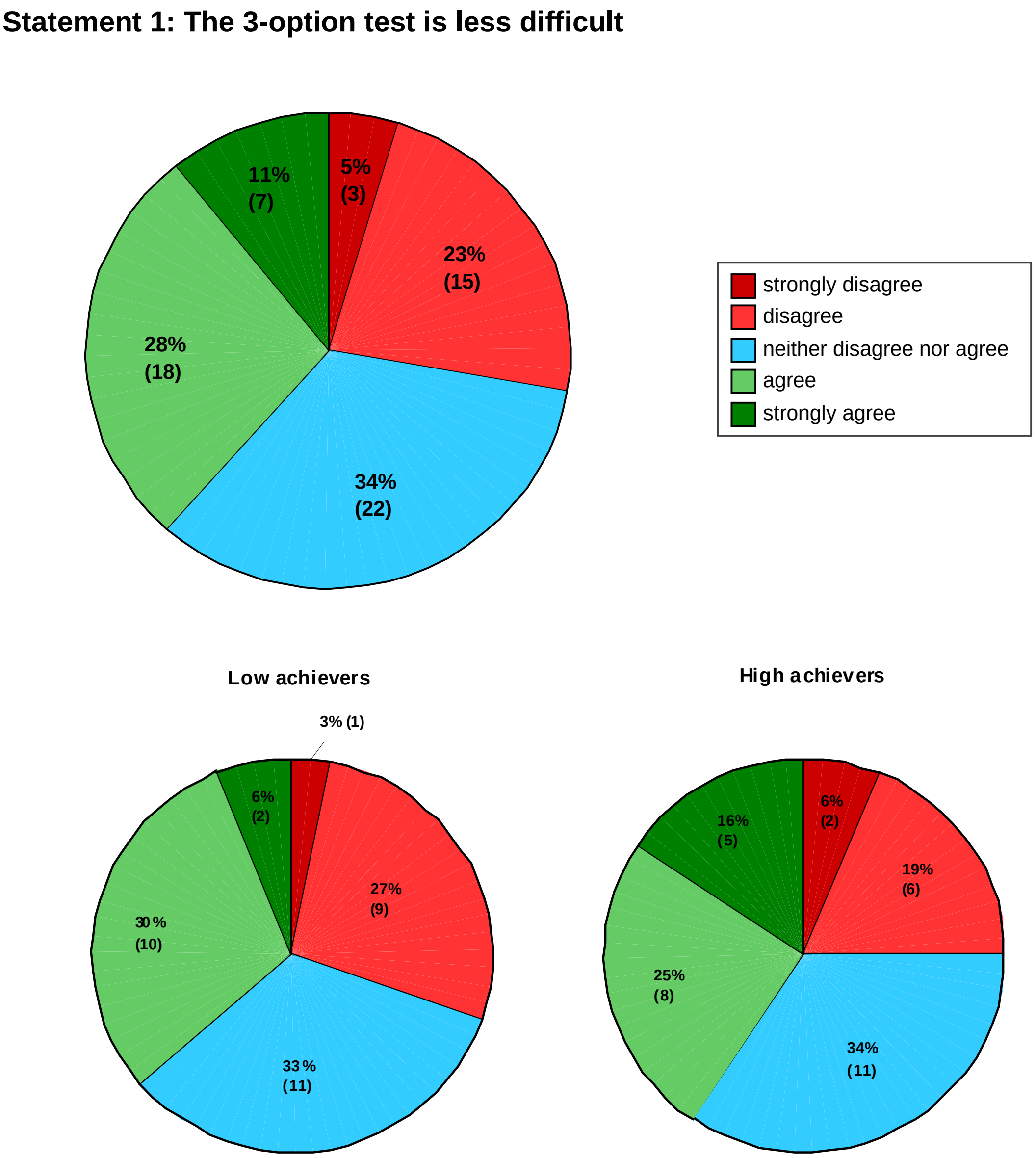
4.3.1 *Perceived difficulty*

First, the differences in perceived difficulty of the 3-option test part as compared to the 4-option test part were examined (see Figure 4.2; exact numbers of respondents per option are given in parentheses). Of particular interest here was whether low achievers perceived the 3-option test differently than the high achievers.

These charts show that the perceptions of the test takers with regard to the difficulty were mixed. A small majority considered the 3-option test as easier than the 4-option test, high achievers more so than low achievers. The low achievers differed most among themselves in their opinions: one third agreed with the statement, one third disagreed, and another 33% neither agreed not disagreed.

In a certain way these results reflect the uncertainty among test takers about whether or not 3-option items are harder than 4-option items. As the data from the actual test performance showed, there were no significant differences between the mean scores in either format, suggesting that one format is not noticeably more difficult than the other. Yet, the 3-option format *appeared* to more than one third of the test takers to be somewhat easier than it actually was, and this may have influenced their overall perception of this format. This is also apparent from the motivations the test takers provided for their preference (see below): while some students indicated that having fewer options makes it more difficult to distinguish clearly wrong answer choices, other test takers thought that this renders the items less difficult.

Figure 4.2: Pie charts of perceived difficulty



Statement 1: The 3-option test is less difficult

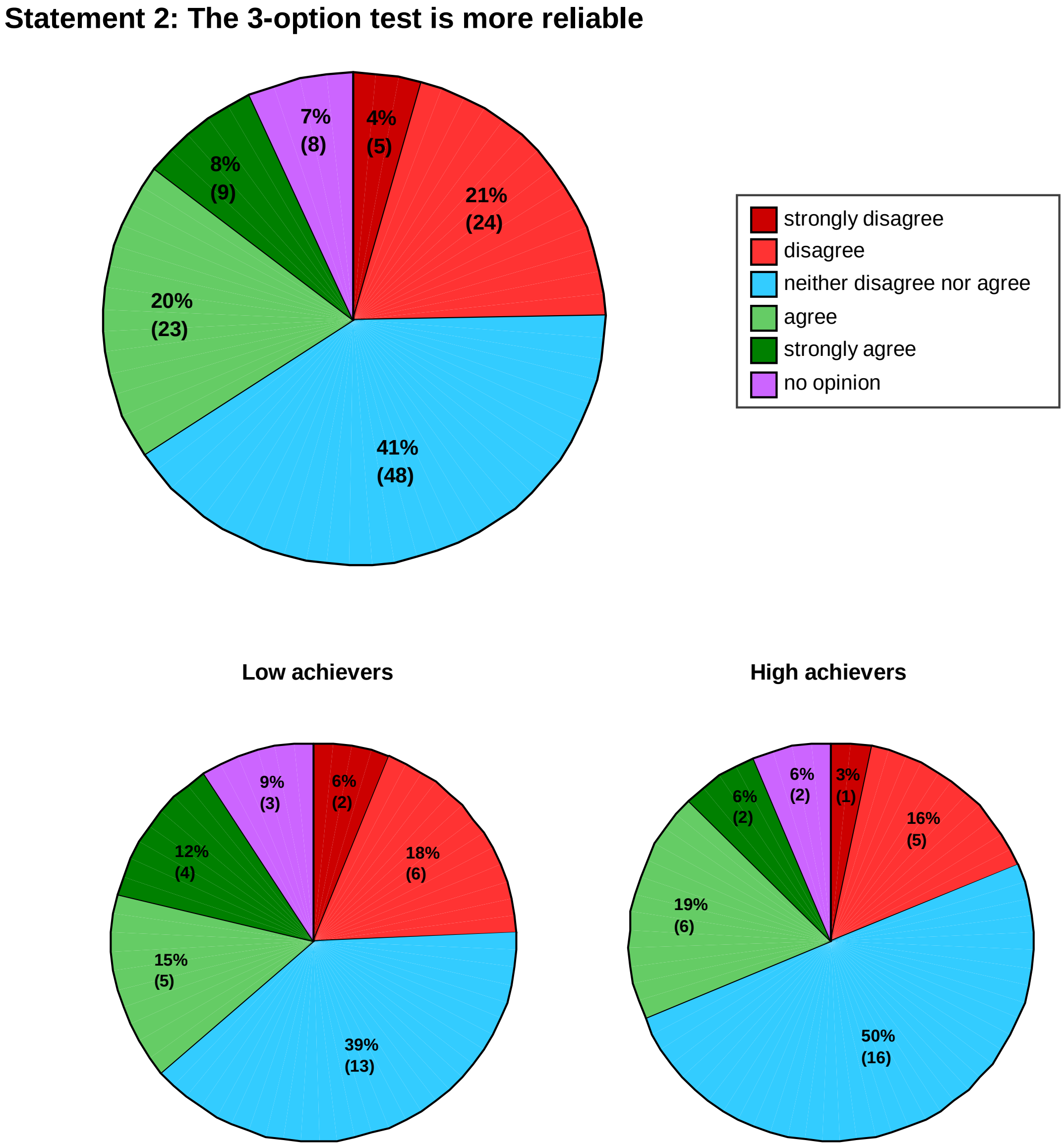
Low achievers

High achievers

4.3.2 *Perceived reliability*

The second topic covered the perceived reliability of the 3-option format in relation to the 4-option format. The questions focussed on topics as the assumed increased probability of getting the answer right by guessing, and whether the 3-option test part measured reading comprehension more or less accurately as the 4-option test part. The frequency of responses per option are presented in Figure 4.3.

Figure 4.3: Pie charts of perceived reliability



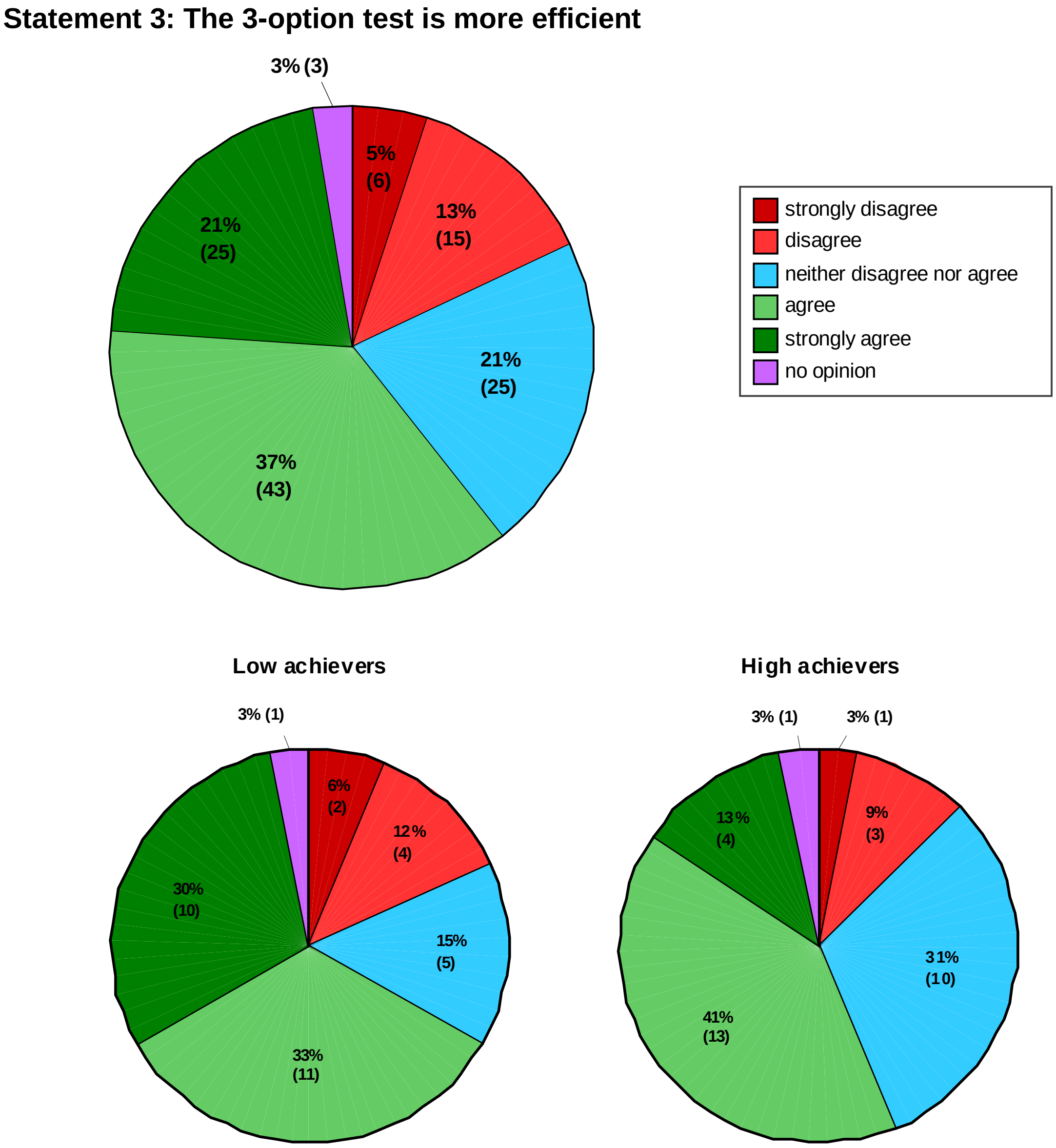
From these charts it can be observed that the majority of test takers did not consider the 3-option format as notably more or less reliable than the 4-option format. Also, there were no great differences between the response patterns of low and high achievers.

However, rather than representing indifference with regard to the statement, the relatively great number of respondents who indicated a neutral position may also indicate unfamiliarity with the notions of reliability and accuracy of measurement. This would explain the somewhat higher number of respondents opting for "don't know" than with the other statements.

4.3.3 *Perceived efficiency*

The third area of interest was the test efficiency: the extent to which test takers appreciated the 3-option format as being time-saving and more practical. The results are displayed in Figure 4.4.

Figure 4.4: Pie charts of perceived efficiency

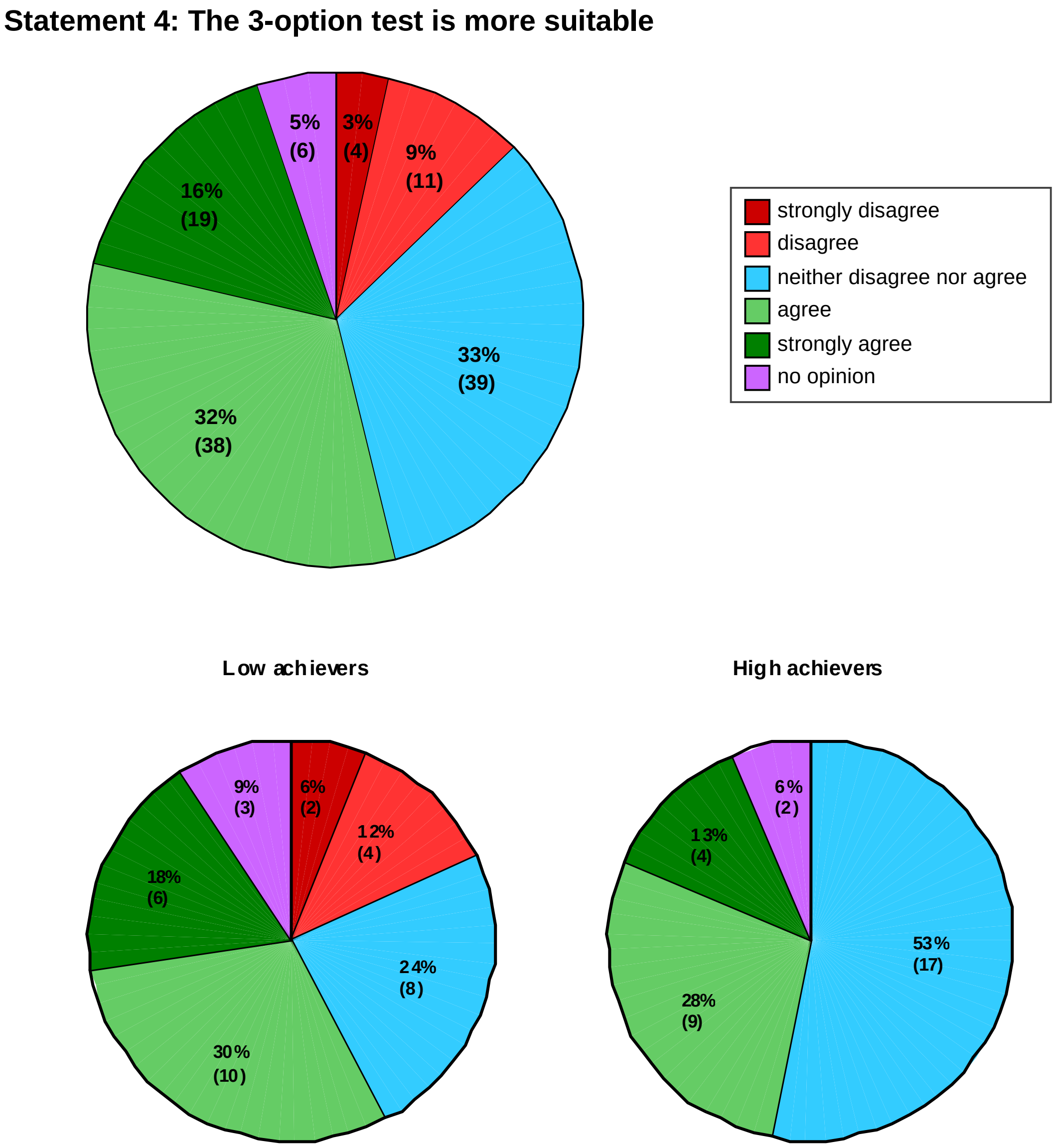


One observation that can be made from these charts is that almost 60% of the respondents considered the 3-option format more efficient than the 4-option format, both in terms of less time needed for responding to an item and being less demanding for their concentration. The gain in efficiency was most appreciated by the low achievers.

4.3.4 *Perceived suitability*

The last subject of investigation was related to the suitability of the 3-option format for testing reading comprehension. The focus was here on the acceptability of the 3-option format, and on the test takers' attitude with regard to whether this format encourages blind guessing. Figure 4.5 shows the answers of the respondents.

Figure 4.5: Pie charts of perceived suitability

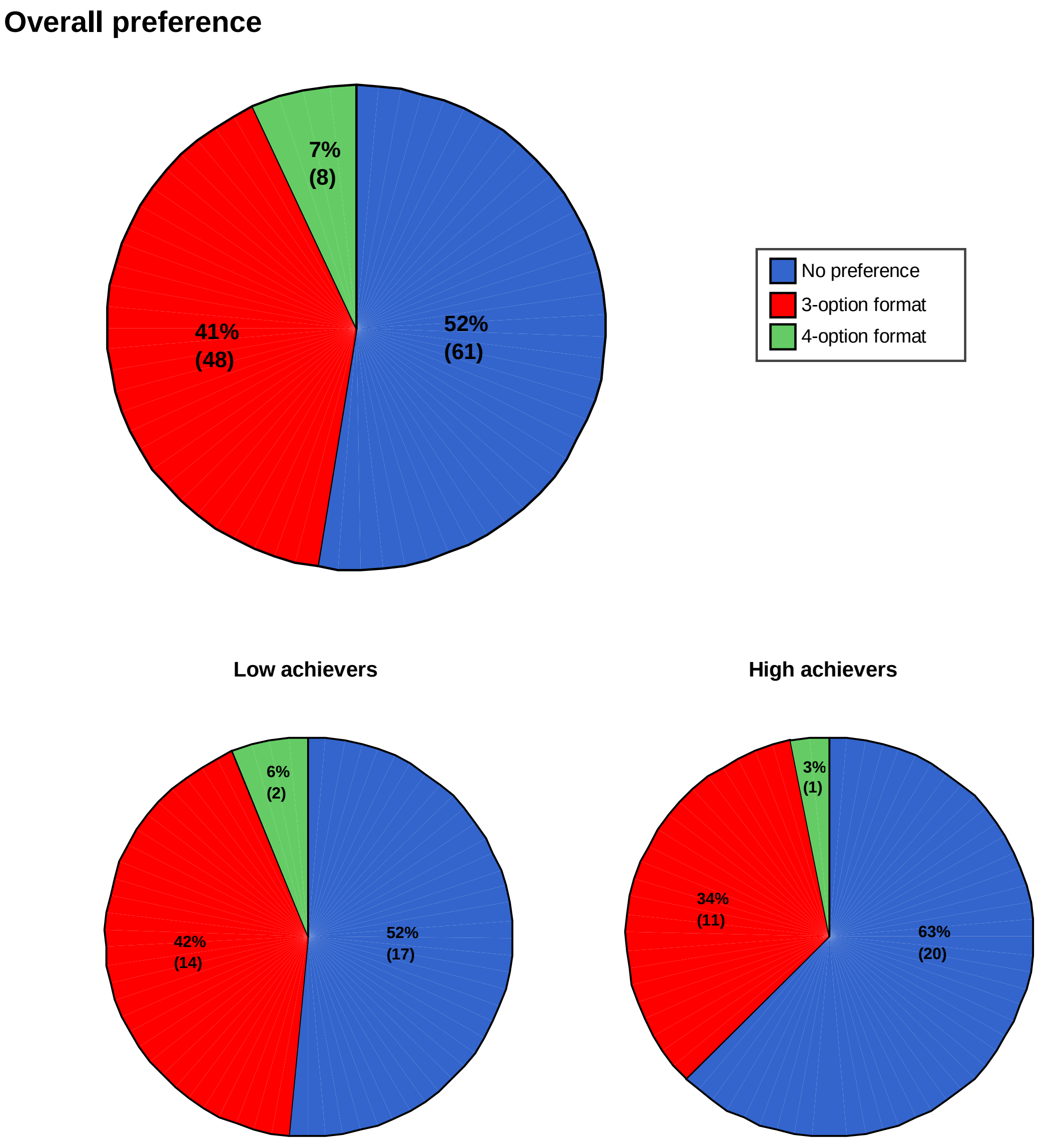


Almost half of all respondents thought that the 3-option format is more suitable than the 4-option format to test reading comprehension. None of the high achievers considered the 3-option test less suitable, but 18% of the low achievers did. One explanation for this could be that, presumably, for the high achievers the issue of guessing was hardly relevant, whereas in the eyes of some of the low achievers blind guessing could possibly provoke undesirable testing behaviour.

4.3.5 *Overall preference*

In the last section of the questionnaire the respondents were asked to state their overall preference for either the 3-option or 4-option format.

Figure 4.6: Pie charts of overall preference



From the responses shown in Figure 4.6 it can be observed that more than 50% of all respondents indicated not to have a preference for one format over the other. Only 8 test takers (7%) preferred the 4-option format, whereas 48 (41%) favoured the 3-option items. The overall preferences of the low-level students were almost identical to the opinions of the entire group of test takers. The high achievers were, if anything, even more outspoken in their indifference regarding the option format: more than 60% had no preference.

In addition to indicating their preference, test takers were asked to give a brief explanation for their choice. In total, 91 respondents (78%) provided a short motivation. The most frequently given answers are summarized in Table 4.13.

Table 4.13: Reasons given for option format preference

Most frequently given explanations	Respondents	
Preference for 3-option format	*N*	*Percentage*
1. Less confusion: less errors due to loss of concentration.	13	34%
2. More efficient: less time needed to respond to an item, therefore more practical and time-saving.	11	29%
3. Easier: less answers to choose from, therefore more chance to get the answer right.	8	21%
4. Other reasons	6	16%
	38	*100%*
Preference for 4-option format		
1. Easier: less hard to detect clearly wrong answer choices.	4	50%
2. More accurate: distinguishes better between those who know the answer and those who do not.	3	38%
3. More fair: less probability to get answer right by guessing.	1	12%
	8	*100%*
No preference		
1. One has to understand the text (reading passage) anyway, regardless of the number of options.	23	51%
2. Equally difficult/reliable/suitable.	17	38%
3. Results are more important than the number of options.	5	11%
	45	*100%*

More than one third of the students favouring the 3-option format mentions as a major advantage that fewer options gives less confusion. It seems then that here some empirical evidence is found for Bruno and Dirkzwager's (1995) theory that having too many options introduces what they call 'noise' (p. 962) into the test item. The additional alternative becomes a "distraction" rather than a distractor, undermining the concentration of low- and high-ability students alike. This overview shows also that, interestingly, both the 3-option and the 4-option format are considered to be easier, although for different reasons.

Finally, a cross tabulation of the responses to the respective statements and the overall preference (see Appendix 5) revealed that there was a fair amount of consistency in the opinions of the test takers. For example, almost 60% of the test takers who neither agreed nor disagreed that the 3-option test was less difficult, more reliable or more suitable expressed no overall preference for either format. Similarly, between 50% and 60% of the respondents who agreed that the 3-option format was less difficult, more reliable and more suitable, favoured the 3-option items. However, of the test takers who thought the 3-option format to be more efficient, the majority had no preference for either 3 or 4 options per item. This suggests that, eventually, the number of options for most test takers was less of a concern than being able to understand the reading passage or in general performing well on the test.

4.3.6 *Summary of results*

Based on the questionnaire responses the following conclusions seem to be justified:

- Low-level test takers did not perceive the relative difficulty and reliability of the 3-option test parts differently than high-level students.
- A majority of the test takers considered the 3-option format more efficient, and at least as suitable as the 4-option format for testing reading comprehension.
- More than 50% of all test takers did not have an explicit preference for either format, and only 8% of the respondents preferred the 4-option items. Generally, the 3-option format was favoured more by the low achievers than by the high achievers.

CHAPTER 5: DISCUSSION

For every problem, there is one solution
which is simple, neat and wrong.
HENRY LOUIS MENCKEN (1880-1956)

5.1 LIMITATIONS

Prior to discussion of the results, limitations of the present study need to be pointed out. First, even though the ERCT was presented to the test takers as an official "trial" examination, it still may have been perceived as a low-stakes test. This may have affected in an undeterminable way their performance on the test and their responses to the questionnaire items. Second, the number of expert judges that participated in this study was limited, and therefore any findings based on their input are vulnerable to sampling error and must be interpreted with caution without further support. Third, it should be emphasized that the 3-option test in this study was created based on 4-option item statistics from previous administrations, using students at generally lower ability levels than those in the present study. As such, it remains unknown (a) whether this has resulted in the removal of distractors which might have been highly discriminating for the sample used in this study, and (b) to what extent the findings are applicable to a situation where such statistics are not available.

Within these limitations, what emerged from the present study were the results summarized below.

5.2 THE RESEARCH QUESTIONS

5.2.1 *Effects on the test quality*

The main purpose of this study was to explore the effect of reducing the number of options on the psychometric properties of the ERCT. The empirically testable criteria on which the 3-option and 4-option formats were compared include item difficulty, item discrimination, internal consistency reliability and efficiency (completion time).

Statistical analyses revealed that the effect of the number-of-options condition on mean item difficulty index, mean point biserial correlation, and test reliability in the four different test parts was nonsignificant. These results are consistent with previous research (e.g., Cizek & O'Day, 1994; Delgado & Prieto, 1998; Shizuka *et al.*, 2006). It was anticipated that the 3-option items would be somewhat easier; presumably, a certain percentage of test takers who would select a distractor in a 4-option item would select the correct response in the 3-option format because of higher probabilities of chance success. Nevertheless, item difficulties remained virtually the same, thus providing support for Ebel's (1968) findings that motivated test takers rarely resort to random guessing when they have sufficient time and the difficulty level is appropriate. It is more likely that instead they choose on the basis of cues derived from the items themselves, irrespective of the number of options provided.

Distractor analysis revealed another important reason why the psychometric properties of the test were not significantly affected by the number of options: only 17% of the 4-option items had 3 effectively functioning distractors and in most cases the fourth option did not contribute at all to the discrimination of the item. In practice, the 4-option test functioned as a 3-option test. These results are consistent with findings by Haladyna and Downing (1993); they found even less items (1-8%) with 2 or 3 effective distractors, but applied somewhat more stringent criteria. However, the present study did not lend support for their finding that the number of effective distractors was unrelated to item difficulty. On the contrary, closer inspection of the data revealed considerable differences in distractor effectiveness between the performance of high- and low-ability students. Whereas the 3-option format appeared to be more efficient for the high achievers, for the low achievers the mean number of discriminating distractors per item was much higher in the 4-option test parts. In the 4-option test the actual responses per item of the low achievers spread over a larger range (3.05 options) than in the 3-option test (2.58 options). This suggests that the elimination of alternatives has greater effect in accordance with the student's ability or, from a different perspective, with the item's difficulty. The less able the student, or the more difficult an item, the greater the spread of choices and therefore the more impact reducing the number of options is likely to have on the information function. These results, then, seem to fit findings by Lord (1977) and Levine and Drasgow (1983) that high-ability test takers may be less inclined to guess, thereby not needing as many options as low-level students who are more inclined to guess. The results in the present study seem to support the hypothesis that information is maximized, and the risk of overestimating achievement is minimized, by using more options per item for lower ability groups and using more items with fewer options for higher ability groups.

With regard to test efficiency, the results in this study concur with previous research, suggesting that time to completion is positively related to the number of options per item. Statistical analysis revealed that the completion time of one 3-option test part was significantly (*p*<.05) shorter than the other test parts. The average completion time of the 3-option test parts was approximately 9% shorter than the 4-option test parts. For the 60-item ERCT this translates to an additional 5 items that could be squeezed in using 3-option items and keeping testing time constant. It remains to be seen whether this number of additional items will lead to a substantial gain in reliability or content validity, which at any rate must be traded off against a loss in extra development time.

Overall, reducing the number of options per item in the kind of comprehension test as the ERCT resulted in a less noticeable shortening of the completion time than in Owen and Froman's (1987) study, who reported a 17% reduction. At the same time, the present study did not find conclusive evidence to support the validity of the assumption by Straton and Catts (1980: 364) that when item stems require long reading times relative to answering the question, the use of 4- or 5-option items would be more desirable. All one could confidently state here is that in such cases, rather than making the 4- or 5-option format more desirable, the benefits of the 3-option format in terms of efficiency are less evident.

It appears, then, that the effect of reducing the number of options on the test efficiency is not such a straightforward matter as it has been presented in previous studies. The time-savings are not merely a function of the number of options, but depend also on the topic and the design of the test. Rogers and Harley (1999) already found that there was no gain in time in the case of a mathematics test where complex computations had to be performed. It seems reasonable to assume that the time-savings are greatest for tests (a) where comparatively most time is spent on processing the options, and (b) which consist of a large number of items, because the time benefits cumulate over more items. The specific design of the ERCT, using only one MC question per reading passage, is such that the time needed to complete each item is spent mainly on reading the passage and proportionally less on answering the question. This may have accounted for the comparatively modest reduction in observed completion times.

5.2.2 *Judging distractor effectiveness*

The second purpose of this study was to examine how reliably test writers can intuitively predict which distractor of 4-option items will be chosen least frequently by test takers. The reliability of their judgement is of crucial importance when instead of dropping one distractor from a 4-option test, a 3-option test will be developed from scratch. Obviously, any savings in development time using 3-option items would be lost if the only way to create a reliable and valid 3-option test is by using 4-option item statistics.

Data collected from 14 item reviewers were analysed to determine the extent to which judgementally deemed least attractive distractors agreed with judgements that were made based on the actual statistical performance of the items. Results showed that – in their combined judgements – subject matter experts exhibited moderately high ability (κ=.57) to identify least functioning distractors when the criterion constituted empirical item analysis data. The probability of making successful predictions appeared to increase if an item had only one distinct least frequently endorsed distractor. The analyses showed further that the combined ratings were on average much more reliable than those by the individual judges. The sample experts (language teachers) were more accurate in their judgements than either the content experts (item writers) or lay judges. Apparently, familiarity with the intended test population enhances the trustworthiness of the judgements. In the present study, the item reviewers made their judgements independently; one may hypothesize that identifying a least frequently chosen distractor by *consensus* of the subject matter experts might result in an even higher agreement with the empirical data.

5.2.3 *Test takers' perceptions and preferences*

A final aim of this study was to address the – in nearly all studies neglected – issue of test takers' perceptions and preferences with regard to the 3-option format. If test takers had a negative attitude towards 3-option items or perceived it as less valid, it might inhibit the face validity, the acceptance, and ultimately the application of this test format, notwithstanding its actual benefits.

The responses to a post-test questionnaire revealed that a majority of the test takers considered the 3-option format more efficient, and at least as acceptable as the 4-option format for testing reading comprehension. Low-level test takers did perceive the relative difficulty and reliability of the 3-option test parts not differently than high-ability students. Most importantly, 61 out of 117 respondents (52.1%) did not have a preference for either format, whereas 41% preferred the 3-option format. These results do not corroborate the findings by Owen and Froman (1987), the only ones so far to examine student preference of option format. They reported that 97.4% of the test takers voted for the 3-option format, 2.6% had no preference, and none chose the (in this case) 5-option format. We can only speculate about the cause of this discrepancy with the results found here, but in contrast with the procedure followed in the present study, Owen and Froman provided their test takers with a summary of the outcomes of the study before asking them to vote for their preferred format. The exact contents of this summary were not given, but possibly, in particular if it highlighted in some way the benefits of the 3-option format, this may have biased the perception of the test takers.

At any rate, the results from the current study are reassuring in the sense that, apparently, most test takers are not guided in their preference by opportunistic motives, such as having a greater probability to get a high score in one format or the other. From the motivations given by the respondents it becomes clear that many test takers acknowledge the fact that, eventually, the number of options is not a decisive factor in answering a question correctly.

CHAPTER 6: CONCLUSIONS

If we knew what we were doing,
it wouldn't be called research, would it?
ALBERT EINSTEIN (1879-1955)

6.1 REVIEW OF THE FINDINGS

The primary impetus behind the present study was the desire to explore whether a 4-option reading comprehension test functions equally well if one option per item would be removed. The results indicate that this is the case: reducing the number of options had no statistically significant effect on the performance of the test takers, irrespective of their ability level. The current study corroborated previous findings suggesting that, from an empirical point of view, the 3-option format is at least as defensible as its 4-option counterpart.

There are many advantages with the use of the 3-option item, three of which are demonstrated in this study. First, administration time is lower with tests containing fewer options. Theoretically, with reduced completion time per item, additional items can be added thereby increasing sampling of content and improving test score reliability. It should be pointed out, however, that so far this argument has never been empirically tested. Second, the number of highly discriminating items increases when the number of options decreases. In most 4-option items the fourth option does not contribute at all to the effectiveness of the item, and it has even been suggested that 3 options per item may be, under most circumstances, a 'natural limit for multiple-choice item writers' (Haladyna & Downing, 1993: 1008). Finally, test takers generally prefer the 3-option format to the 4-option format, which enhances the face validity and acceptability of MC questions consisting of only 3 options.

Other obvious benefits from using the 3-option format are:

- item writing is less laborious. In the context of test development, there is always a need to develop new items, and this effort can be lessened when 3 instead of 4 options are used; since the fourth – often implausible – option typically takes more time to come up with than the other three, the amount of time and energy saved is relatively greatest;

- the chances of providing unintended cues that profit test-wise students will be decreased (Owen & Froman, 1987). Limiting the number of distractors will reduce the likelihood of including distractors which may weaken the effectiveness of the item; for example, distractors that are similar to existing distractors, specific determiners, or absurd distractors;
- students can answer questions with less distractions; using fewer options will reduce both memory load and the possible confusion of thought resulting from perusal of several wrong answers; students might less quickly get lost reading many alternatives and have to return and reread the question and early alternatives;
- students will feel less pressured because they can work more slowly or spend time to recheck; and
- the distractors taken as a set should be more plausible.

Of course, 3 options are not in all circumstances better than 4. If item analysis demonstrates that some items have 4 highly effective options, it makes little sense to undo good construction by discarding useful information. But, generally, using more options does little to improve item and test score statistics and typically results in implausible distractors. Item writers sometimes solve this problem by adding inclusive options such as *all of the above* or *none of these*, the use of which should be discouraged on logical as well as practical grounds: they seem to draw examinees into test-taking strategies instead of actually testing student knowledge (Crehan *et al.*, 1993).

6.2 IMPLICATIONS

The findings of this study have useful implications for test development using the MC format. First, additional evidence has been gathered to support the fairly consistent body of research indicating that the use of 3 options is optimal for MC items. In this study, items with only 3 options performed as well as – and, by some measures, more effectively than – the same items with 4 options. These findings also pointed to some potential practical application that may be confidently implemented if additional research involving other content areas supports the findings. First, the results suggest that 4-option MC items can be safely reduced to 3-option items by removing a least frequently chosen distractor. If generalisable, this finding may be of some comfort to testing programs currently using 4-option items and desiring to move to the 3-option format by removing non-functioning options from existing items. The findings of this study also suggest that, whenever item analysis data are unavailable, expert identification of least functioning distractors is fairly reliable, provided that the selection is made on the basis of combined ratings of several judges. Taken one step further, these results provide tentative support for the assumption that item writers could rely largely on their intuition and expertise in deciding which distractor would be potentially non-functioning, making it possible to develop a 3-option test from the beginning that is equally reliable and discriminating as a 3-option test created based on 4-option item statistics. However, as this study indicated, these item writers would do well to use (preferably a group of) language teachers familiar with the intended test population for reviewing the plausibility of their distractors. Clearly, research into whether this hypothesized result would actually be obtained should be conducted before such a procedure is implemented.

6.3 FUTURE DIRECTIONS

One issue that seemingly refutes the optimality of the 3-option format is the finding in this study that for low-ability students information is maximized, and the risk of overestimating achievement is minimized, by using 4 options or more. The problem here is that "high-ability" and "low-ability" are relative qualifications: a student who is a low achiever on one test may be a high achiever on another. Thus, maximizing the information per test would require the option format to change according to the test taker's ability level with respect to a particular test. In most testing situations this may be difficult to realize. However, a solution might be found in the application of computer-adaptive testing software. It is technologically conceivable that a computer programme, after having determined the overall ability level of the student with a few items, automatically presents the student with the optimum number of options per item, depending on the test taker's ability level. High-level students would thus get more items with fewer options, and low-level students less items with more options, thus rendering a computer-adaptive test even more "adaptive". This would certainly be a topic that merits more research because it would make it possible to maximize information for each test taker individually.

Computerized testing would also provide an excellent means to investigate another area of interest more closely: the effects of the number of options on the test completion time. In the present study, data were collected on the completion times of the two test parts only, because it was thought that asking test takers to provide more detailed data (for example time needed per proficiency level, or even per item) would be too disrupting for them. A computer programme, on the other hand, generates such data automatically and without interference. More detailed information would give better insight in, for instance, how much longer it takes to respond to a Level 3 item compared to a Level 1 item, or the difference in exact responding time per item between high achievers and low achievers.

Further, given the optimality of the 3-option format, more research is warranted to examine whether and to what extent two types of 3-option tests differ in terms of psychometric characteristics: one produced as a 3-option test from the beginning and the other created by dropping one distractor of a 4-option test.

Finally, it would be valuable to find out whether the results obtained in the present study may be applicable to a *listening* comprehension test. In a listening test it is important that the item should demand as little effort of understanding as possible from the test takers, because the intention is to assess their understanding of what they hear (the passage), not of what they read on their test sheet (the items). Another major consideration that distinguishes listening tests from reading tests involves memory. It might be assumed that the risk of memory overload is directly related to the number of options per item. Thus, in a listening test the number of options might be a more critical factor than in a reading comprehension test. As such, this issue must be considered an urgent matter for further investigation.

REFERENCES

Abad, F.J., Olea, J. and Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema*, 13 (1): 152-158.

Alderson, J.C. (1993). Judgments in language testing. In D. Douglas and C. Chapelle (eds.), *A new decade of language testing research.* Alexandria, VA: TESOL, pp. 46-57.

Alderson, J.C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Banerjee, J. (2004). Qualitative analysis methods. Section D of The Reference Supplement to the Preliminary Pilot version of the Manual for *Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.* Strasbourg: Language Policy Division.

Berríos, G., Rojas, C., Cartaya, N. and Casart, Y. (2005). Effect of the number of options on the quality of est reading comprehension multiple-choice exams. *Paradìgma*, 26 (1): 89-116.

Brown, J.D. (2005). *Understanding Research in Second Language Learning.* Cambridge: Cambridge University Press.

Bruno, J.E. and Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement*, 55 (6): 959-966.

Budescu, D.V. and Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula-scoring. *Journal of Educational Measurement*, 30 (4): 277-291.

Budescu, D.V. and Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement*, 22 (3): 183-196.

Bussis, A. and Chittenden, E. (1987). Research currents: What the reading tests neglect. *Language Arts*, 64 (3): 302-308.

Cizek, G.J. and O'Day, D.M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement*, 54 (4): 861-872.

Cizek, G.J., Robinson, K.L. and O'Day, D.M. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement*, 58 (4): 605-611.

Cohen, L., Manion, L. and Morrison, K. (2007). *Research Methods in Education*. (Sixth edition). London: Routledge.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement*, 30: 353-358.

Costin, F. (1972). Three-choice versus four-choice items: implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement*, 32: 1035-1038.

Costin, F. (1976). Difficulty and homogeneity of three-choice versus four-choice objective test items when matched for content of stem. *Teaching of Psychology*, 3 (3): 144-145.

Crehan, K.D., Haladyna, T.M. and Brewer, B.W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, 53: 241-247.

Delgado, A.R. and Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, 14 (3): 197-201.

Dőrnyei, Z. ((2003). *Questionnaires in Second Language Research. Construction, Administration, and Processing*. Mahwah, NJ: Erlbaum.

Ebel, R.L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement*, 29: 565-570.

Ebel, R.L. (1972). *Essentials of educational measurement.* (Second edition). Englewood Cliffs, NJ: Prentice-Hall.

Farr, R., Pritchard, R. and Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27 (3): 209-226.

Frary, R.B. (1980). The Effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, 4 (1): 79-90.

Green, K., Sax, G. and Michael, W.B. (1982). Validity and reliability of tests having differing numbers of options for students of differing levels of ability. *Educational and Psychological Measurement*, 42 (1): 239-245.

Grier, J. Brown (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement*, 12 (2): 109-113.

Grier, J. Brown (1976). The optimal number of alternatives at a choice point with travel time considered. *Journal of Mathematical Psychology*, 14: 91-97.

Haladyna, T.M. (2004). *Developing and Validating Multiple-Choice Test Items.* (Third edition). Mahwah, NJ: Erlbaum.

Haladyna, T.M. and Downing, S.M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2 (1): 37-50.

Haladyna, T.M. and Downing, S.M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53: 999-1010.

Haladyna, T.M., Downing, S.M. and Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15 (3): 309-334.

Henning, G. (1987). *A Guide to Language Testing. Development, Evaluation, Research.* Cambridge, MA: Newbury House.

Hopkins, K.D. (1998). *Educational and Psychological Measurement and Evaluation.* (Eighth edition). Needham Heights, MA: Allyn and Bacon.

Lado, R. (1965). *Language Testing.* London: Longmans.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33: 159-174.

Landrum, R.E., Cashin, J.R. and Theis, K.S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement*, 53: 771-778.

Levine, M.V. and Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement*, 43: 675-685.

Lord, F.M. (1977). Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement*, 14 (1): 33-38.

LTEST-L, Language Testing Research and Practice. Online discussion forum (mailing list), LTEST-L@LISTS.PSU.EDU.

North Atlantic Treaty Organization (2003): *NATO Standardization Agreement* (STANAG) *6001*: Language Proficiency Levels (Edition 2). Retrieved 20 November 2003, from http://www.dlielc.org/bilc/reports_1.html.

Owen, S.V. and Froman, R.D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement*, 47: 513-522.

Ramos, R.A. and Stern, J. (1973). Item behavior associated with changes in the number of alternatives in multiple choice items. *Journal of Educational Measurement*, 10 (4): 305-310.

Rogers, W.T. and Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, 59 (2): 234-247.

Rogers, W.T. and Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, 12 (3): 247-259.

Shizuka, T., Takeuchi, O., Yashima, T. and Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, 23 (1): 35-57.

Sidick, J.T., Barrett, G.V. and Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology*, 47: 829-835.

Straton, R.G. and Catts, R.M. (1980). A comparison of two, three and four-choice items tests given a fixed total number of choices. *Educational and Psychological Measurement*, 40: 357-365.

Trevisan, M.S., Sax, G. and Michael, W.B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, 51 (4): 829-837.

Trevisan, M.S., Sax, G. and Michael, W.B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement*, 54 (1): 86-91.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, 1: 386-391.

Weitzman, R.A. (1970). Ideal Multiple-Choice Items. *Journal of the American Statistical Association*, 65 (329): 71- 89.

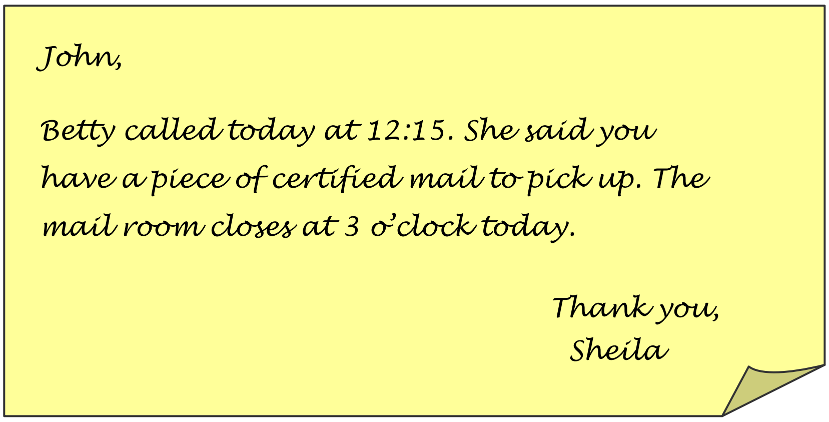
APPENDICES

APPENDIX 1

Illustrative samples of test items at the various proficiency levels

Sample Level 1 test item

A message at the office



John,

Betty called today at 12:15. She said you have a piece of certified mail to pick up. The mail room closes at 3 o'clock today.

Thank you,
Sheila

This note tells John to

A. close the mail room at three.
B. go to get some mail. *
C. mail a letter for Betty.
D. pick up Betty at the mail room.

Sample Level 2 test item

A news item:

> South Africa is shooting pigeons in its diamond producing area because the birds are being used to smuggle gems out of the country. Diamonds are leaving the country in an extremely worrisome manner: strapped onto the bodies of pigeons and flown out of the country. The law is now to shoot all pigeons on sight. Mineworkers have been implicated in the widespread theft, and diamond producers will need to spend about $8 million to improve security.

Pigeons are in the news because they are

A. part of a plan to prevent diamond smuggling.
B. part of a safety program for mineworkers.
C. being shot to prevent spread of a disease.
D. being used in a criminal activity.*

* = key

Sample Level 3 test item

An editorial

POLAR BEAR A SYMBOL OF GLOBAL WARMING

The U.S. is nearly a month overdue in making a decision on whether to list the polar bear as a threatened species. Though there's reason to view the delay with cynicism – it gave the government time to lease prime polar-bear habitat for oil exploration – this is a delay with far-reaching and potentially unintended consequences.

The polar bear would become the first species listed as a result of global warming rather than direct causes, such as construction in critical habitat, hunting or exposure to toxic substances.

Beyond that, the bear doesn't appear threatened at first glance. There's no evidence that the bear's numbers are declining, the usual trigger for listing a species. It certainly wouldn't qualify as "endangered" – on the brink of extinction.

"Threatened" is another matter, requiring only a finding that if conditions don't change, an animal is in danger of eventually sliding toward extinction. For this, the evidence is solid. Polar bears spend much of their time not on land but on ice floes, where they hunt and raise their young. The ice has been melting, and polar bears are showing signs of distress as they make longer swims. Three years ago, scientists found that some bears had probably drowned after swimming long distances, unable to find a nearby sheet of ice. At the current rate, the prediction is that 80 percent of the summertime ice floes will disappear within 20 years.

But if the reasons behind the polar bear's possible inclusion on the threatened list are indirect and complex, so are many of the possible ramifications. Drilling for oil in the bear's hunting waters would appear an obvious problem. But what about the motorists, thousands of miles away, using that oil to drive to work, emitting greenhouse gases as they go? To put it straightforwardly, simply being human and alive contributes to carbon emissions.

The question of how far to go to protect the polar bear quickly becomes a debate about how much we should change our habits to slow the pace of climate change. Reports of diminished glaciers and shifting weather patterns haven't grabbed the public's imagination. A snowy white bear is another matter. The polar bear gives us a tangible reason to recognize that global warming is real and that it matters. Let the conversation begin.

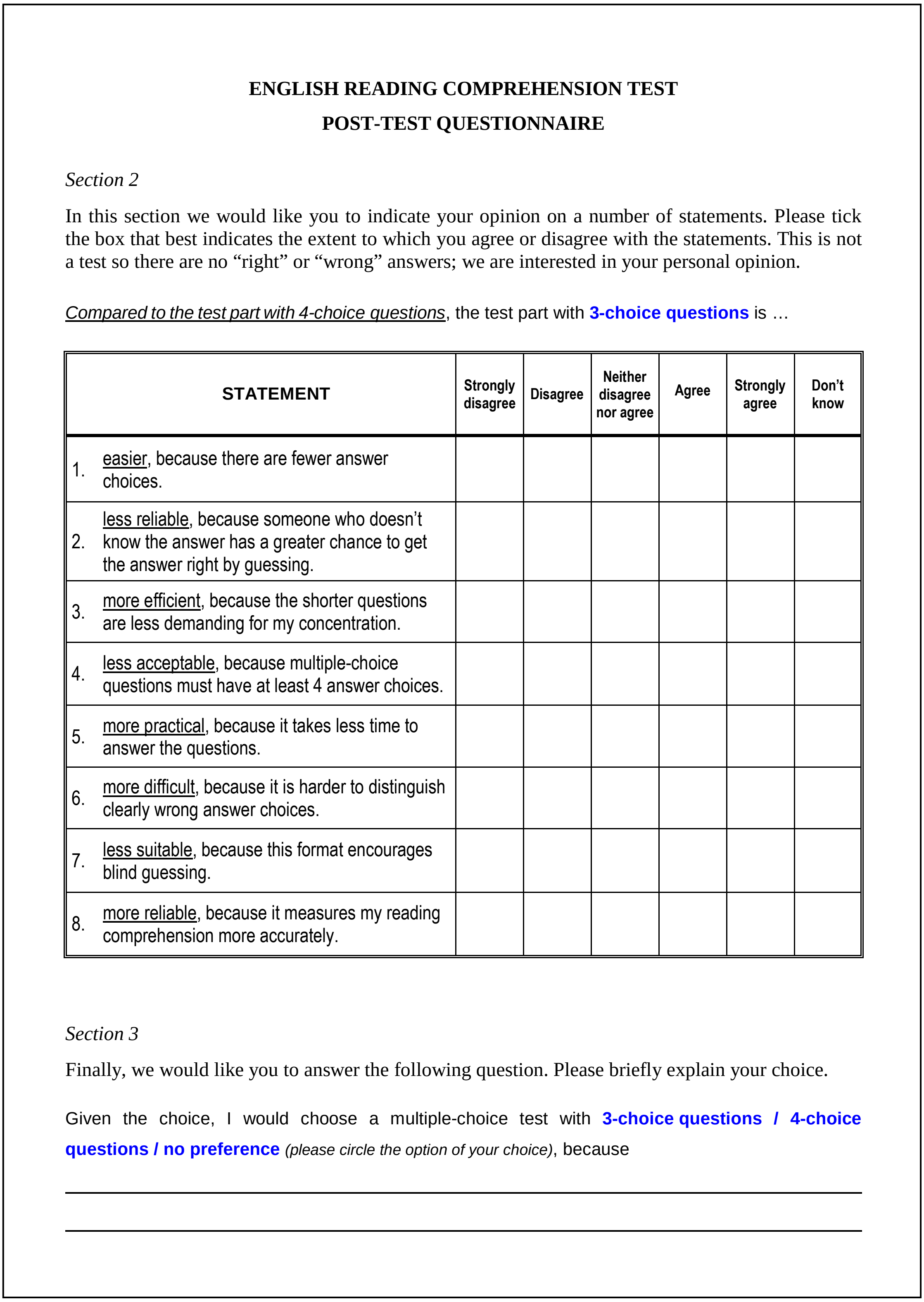
This writer makes the point that

A. the polar bear's plight is directly related to oil drilling within their habitat.
B. the public has begun to express concerns about shifting weather patterns.
C. polar bears qualify for "endangered" status because of probable drownings.
D. changing everyday behaviour is a critical factor in preventing global warming.*

* = key

APPENDIX 2

Post-test Questionnaire *(partial)*



ENGLISH READING COMPREHENSION TEST

POST-TEST QUESTIONNAIRE

Section 2

In this section we would like you to indicate your opinion on a number of statements. Please tick the box that best indicates the extent to which you agree or disagree with the statements. This is not a test so there are no “right” or “wrong” answers; we are interested in your personal opinion.

Compared to the test part with 4-choice questions, the test part with **3-choice questions** is …

STATEMENT	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly agree	Don’t know
1. easier, because there are fewer answer choices.						
2. less reliable, because someone who doesn’t know the answer has a greater chance to get the answer right by guessing.						
3. more efficient, because the shorter questions are less demanding for my concentration.						
4. less acceptable, because multiple-choice questions must have at least 4 answer choices.						
5. more practical, because it takes less time to answer the questions.						
6. more difficult, because it is harder to distinguish clearly wrong answer choices.						
7. less suitable, because this format encourages blind guessing.						
8. more reliable, because it measures my reading comprehension more accurately.						

Section 3

Finally, we would like you to answer the following question. Please briefly explain your choice.

Given the choice, I would choose a multiple-choice test with **3-choice questions / 4-choice questions / no preference** *(please circle the option of your choice)*, because

____

____

APPENDIX 3

Descriptive and referential statistics parallel tests

Table A3.1: Descriptive statistics Form A and Form B

Statistic	Form A	Form B
Valid *N*	62	62
Missing *N*	0	0
No. of items	60	60
Max. score	60	60
Mean	47.26	47.60
Median	48.50	48.00
Mode	51	48; 51
Std. deviation	6.909	6.637
Mean item difficulty	.78	.79
Std. Deviation *p* values	.36	.36
Variance	47.736	44.048
Skewness	-.433	-.828
Std. error of skewness	.304	.304
Kurtosis	-.565	1.092
Std. error of kurtosis	.599	.599
Range	28	33
Minimum	30	27
Maximum	58	60

Table A3.2: Reliability Statistics Form A and Form B

Form A

Each of the following component variables has zero variance and is removed from the scale: item 2, item 38

The determinant of the covariance matrix is zero or approximately zero. Statistics based on its inverse matrix cannot be computed and they are displayed as system missing values.

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.836	.834	58

Form B

Each of the following component variables has zero variance and is removed from the scale: item 32, item 40

The determinant of the covariance matrix is zero or approximately zero. Statistics based on its inverse matrix cannot be computed and they are displayed as system missing values.

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.822	.845	58

Table A3.3: Descriptive statistics Anchor Items Group 1 and Group 2

Statistic	**Group 1**	**Group 2**
Valid *N*	62	62
Missing *N*	0	0
No. of items	6	6
Max. score	6	6
Mean	4.52	4.71
Median	5.00	5.00
Mode	5	6
Std. deviation	1.112	1.233
Variance	1.237	1.521
Skewness	-.190	-.503
Std. error of skewness	.304	.304
Kurtosis	-1.042	-1.011
Std. error of kurtosis	.599	.599
Range	4	4
Minimum	2	2
Maximum	6	6

Table A3.4: Correlations Test Parts Form A and Form B

Correlation Form A

		Score Part 1 (4 options)	Score Part 4 (3 options)
Score Part 1 (4 options)	Pearson Correlation	1	.752(**)
	Sig. (2-tailed)		.000
	N	62	62
Score Part 4 (3 options)	Pearson Correlation	.752(**)	1
	Sig. (2-tailed)	.000	
	N	62	62

** Correlation is significant at the 0.01 level (2-tailed).

Correlation Form B

		Score Part 2 (4 options)	Score Part 3 (3 options)
Score Part 2 (4 options)	Pearson Correlation	1	.688(**)
	Sig. (2-tailed)		.000
	N	62	62
Score Part 3 (3 options)	Pearson Correlation	.688(**)	1
	Sig. (2-tailed)	.000	
	N	62	62

** Correlation is significant at the 0.01 level (2-tailed).

APPENDIX 4

Correction for scoring

The question of whether or not to correct scores for guessing is a recurring issue in MC testing. Theoretically, the probability of getting an item right by chance is larger for 3-option items than for 4-option items (33% against 25%). The most common method of formula scoring levies a penalty of 1/(*k*-1) points against each incorrect answer, yielding a corrected score of *S' = R–W/(k–1)*, in which *R* stands for the number of items answered rightly, *W* for the number of questions answered wrongly, and *k* for the number of options per item.

While acknowledging that the increased chance probability to get an item right could lead to an increase in performance on the 3-option items, for several reasons it was decided not to correct the scores for guessing in this investigation: first, the test takers participating in this study had been encouraged to answer all items even if they were not sure. Formula scoring corrects for guessing by penalizing incorrect responses, while being neutral regarding omitted items; therefore, applying a correction for guessing would have been unfair and invalid in this case. Further, among measurement experts there is considerable controversy about formula scoring (*cf.*, e.g., Lado, 1965: 367; Frary, 1980; Budescu & Bar-Hillel, 1993). Correction for scoring is often criticized on the ground that it is based on a false assumption – the assumption that all correct answers are the result of knowledge and that all wrong answers are guessed wrong. Because of the invalidity of this assumption underlying the formula, and because scores corrected for guessing tend to include 'irrelevant measures of the test taker's testwiseness or willingness to gamble' (Ebel, 1972: 256), the use of the formula is not generally recommended. A final consideration in the decision not to apply formula scoring in this study were the earlier mentioned research findings suggesting that motivated test takers rarely resort to blind guessing, and that guessing generally has a negligible effect on the test score.

APPENDIX 5

Cross tabulation of overall preferences and perceived properties

		The 3-option format is less difficult					Total	The 3-option format is more reliable						Total	The 3-option format is more efficient						Total	The 3-option format is more suitable						Total
Overall preference		strongly disagree	dis-agree	neither disagree nor agree	agree	strongly agree		strongly disagree	dis-agree	neither disagree nor agree	agree	strongly agree	no opinion		strongly disagree	dis-agree	neither disagree nor agree	agree	strongly agree	no opinion		strongly disagree	dis-agree	neither disagree nor agree	agree	strongly agree	no opinion	
No preference	Count	3	13	31	12	2	61	2	11	28	7	8	5	61	2	7	12	24	14	2	61	2	4	22	18	12	3	61
	% within Overall preference	4.9%	21.3%	50.8%	19.7%	3.3%	100%	3.3%	18.0%	45.9%	11.5%	13.1%	8.2%	100%	3.3%	11.5%	19.7%	39.3%	23.0%	3.3%	100%	3.3%	6.6%	36.1%	29.5%	19.7%	4.9%	100%
	% within Perceived property	50.0%	72.2%	59.6%	36.4%	25.0%	52.1%	40.0%	45.8%	58.3%	30.4%	88.9%	62.5%	52.1%	33.3%	46.7%	48.0%	55.8%	56.0%	66.7%	52.1%	50.0%	36.4%	56.4%	47.4%	63.2%	50.0%	52.1%
3-option format	Count	1	5	17	19	6	48	2	10	19	13	1	3	48	3	6	11	19	8	1	48	1	6	14	19	5	3	48
	% within Overall preference	2.1%	10.4%	35.4%	39.6%	12.5%	100%	4.2%	20.8%	39.6%	27.1%	2.1%	6.3%	100%	6.3%	12.5%	22.9%	39.6%	16.7%	2.1%	100%	2.1%	12.5%	29.2%	39.6%	10.4%	6.3%	100%
	% within Perceived property	16.7%	27.8%	32.7%	57.6%	75.0%	41.0%	40.0%	41.7%	39.6%	56.5%	11.1%	37.5%	41.0%	50.0%	40.0%	44.0%	44.2%	32.0%	33.3%	41.0%	25.0%	54.5%	35.9%	50.0%	26.3%	50.0%	41.0%
4-option format	Count	2	0	4	2	0	8	1	3	1	3	0	0	8	1	2	2	0	3	0	8	1	1	3	1	2	0	8
	% within Overall preference	25.0%	.0%	50.0%	25.0%	.0%	100%	12.5%	37.5%	12.5%	37.5%	.0%	.0%	100%	12.5%	25.0%	25.0%	.0%	37.5%	.0%	100%	12.5%	12.5%	37.5%	12.5%	25.0%	.0%	100%
	% within Perceived property	33.3%	.0%	7.7%	6.1%	.0%	6.8%	20.0%	12.5%	2.1%	13.0%	.0%	.0%	6.8%	16.7%	13.3%	8.0%	.0%	12.0%	.0%	6.8%	25.0%	9.1%	7.7%	2.6%	10.5%	.0%	6.8%
Total	Count	6	18	52	33	8	117	5	24	48	23	9	8	117	6	15	25	43	25	3	117	4	11	39	38	19	6	117
	% within Overall preference	5.1%	15.4%	44.4%	28.2%	6.8%	100%	4.3%	20.5%	41.0%	19.7%	7.7%	6.8%	100%	5.1%	12.8%	21.4%	36.8%	21.4%	2.6%	100%	3.4%	9.4%	33.3%	32.5%	16.2%	5.1%	100%
	% within Perceived property	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%