



<https://theses.gla.ac.uk/>

Theses Digitisation:

<https://www.gla.ac.uk/myglasgow/research/enlighten/theses/digitisation/>

This is a digitised version of the original print thesis.

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study,
without prior permission or charge

This work cannot be reproduced or quoted extensively from without first
obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any
format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author,
title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

REASONING ABILITY IN SUBJECTS OF HIGH INTELLIGENCE:

AN EXPERIMENTAL STUDY OF INDIVIDUAL DIFFERENCES

with an Appendix on Valentine's Reasoning Tests
for Higher Levels of Intelligence

by

Ian G. Wallace, M.A., B.Phil., M.Ed.

Submitted for the degree of Ph.D. at the University of Glasgow in 1970

ProQuest Number: 10800640

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10800640

Published by ProQuest LLC (2018). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

PREFACE

In November 1967 the 320 undergraduate members of the Ordinary Psychology class at the University of Glasgow sat the published form and an alternative form of Valentine's Reasoning Tests for Higher Levels of Intelligence on successive Fridays and Terman's Concept Mastery Test on the intervening Tuesday. The purpose to be served by this programme of testing was twofold: to assess the adequacy of the alternative form of the Valentine test, which the present writer had devised, and to obtain information about this large group of undergraduates which would make it possible to constitute a group of 'poor reasoners' and a group of 'good reasoners' matched in respect of general academic ability and attainment.

In the event the alternative form did appear to be adequate in the sense that the means and standard deviations of the scores on the two forms were reasonably close when account was taken of the order in which subjects had sat the two forms. Accordingly, results on both forms - as well as on the Terman test - were used for the second of the purposes mentioned above. In a subsequent, more detailed, analysis of the items of the two forms of the Valentine, however, certain serious faults of detail became apparent and these, together with some aspects of the scoring system which had seemed curious even on a more superficial acquaintance, suggested the usefulness of a more extended appraisal of the original.

The main part of this thesis is closely tied to Section B of the Valentine Reasoning Tests, the section devoted to 'deductive reasoning' problems and the basis of the division, referred to above, of subjects into those of deductive reasoning ability commensurate with their superior academic attainment and ability as measured in other ways and those of comparable general ability whose success on deductive reasoning tasks was very much less. Accordingly, a familiarity with this section of the test (in its two forms), both in the sense of a knowledge of its contents and of an appreciation of its strengths and weaknesses, is, if not essential, at least important for the reader of this thesis. At the same time, the discussion which naturally grew out of the more extended appraisal of the test referred to above, although, I think, of some

importance in its own right in view of the enthusiasm with which it has been received in some influential quarters, goes beyond the limits of what is strictly relevant to the main concern of this thesis. I have therefore relegated it to the first of the Appendices and the reader who is unfamiliar, or insufficiently familiar, with the test may find it useful to look at some of the points made in the first six sections of Appendix A as well as at the two forms of the test itself, reproduced at the beginning of Appendix B.

The Valentine test was intended by its author (1957, 1961) as a means of assessing intelligence at the highest levels and it has been welcomed as such by Anstey (1966, p. 231) who describes it as 'an extremely clever and promising one'. In the last section of Appendix A I have tried to show that there are important theoretical objections to the use of the test for this purpose, while insisting on those special qualities which commended it to me as a test of deductive reasoning ability in the first place. In any case I should like to think that this thesis as a whole will appear to make at least a minor contribution to the marriage of its mental testing and experimental divisions which Butcher (1968, p. 10) has seen as so pressing a need for contemporary psychology.

In the course of completing the work described in this thesis I have received help and advice from a very large number of people to whom I should like to express my sincere gratitude. These include, of course, my colleagues on the teaching staff in the Psychology Department who have borne patiently with me at all times and readily offered advice when this was sought; the secretarial and technical staff of this Department and of the Adam Smith Building more generally who prepared materials at what must often have seemed to them to be unreasonably short notice; former colleagues in the Logic Department, especially Mr. T. Greenwood, Mr. P. Shaw and Mr. E. Toms; Mr. P. Covey-Crump of the Statistics Department; Miss C. Duff and technical assistants in the Computing Department; the Advisers of Studies in Arts and Science and their secretarial staffs for help in locating the academic records of my subjects; the trustees of the William Boyd Prize Fund for a small grant to enable me to pay my subjects for participation in the second stage of my research; the subjects themselves for their patient co-operation and interest in the research in which they played an indispensable part; the publishers, Oliver and Boyd, of Edinburgh for permission to reproduce the published form of the Valentine test; my wife for assistance and

support throughout the period during which this thesis was being prepared; and my supervisor, Professor R. W. Pickford, for his ready help in practical matters, his encouragement and his advice, particularly in matters of broad strategy.

TABLE OF CONTENTS

	Page
Preface	ii
Summary	vii
 Chapter	
1. <u>Introduction and Critical Review of the Literature</u>	1
1.1 Introduction	1
1.2 Some prefatory remarks on the relevant literature	4
1.3 Wilkins on the effect of changing the material in terms of which a reasoning task is presented	5
1.4 The so-called 'atmosphere effect'	13
1.5 The importance of attitudes towards the conclusion of a syllogistic argument	24
1.6 Some more general considerations	34
1.7 Summary of the main points	38
2. <u>Selection of Subjects, Preliminary Evidence, and the 'Five Types of Statements' Task, Part One: Description of the Materials and Procedure</u>	41
2.1 Selection of subjects	41
2.2 Some preliminary evidence	44
2.3 The 'five types of statement' task: rationale and procedure	50
3. <u>The 'Five Types of Statement' Task, Part Two: Results, Analysis and Discussion</u>	61
3.1 Differences in the time taken to complete the task	62
3.2 Difficulty measured in terms of errors	66
3.3 Two more general points of comparison between the groups	86
3.4 The nature of the experimental task: a discussion with reference to the work of Wason and others	89
3.5 The fifth series of trials	101
3.6 Review of the main points	107
4. <u>The 'Four Types of Statement' Task</u>	111
4.1 Introduction	111
4.2 The subjects	113
4.3 The 4TS materials	113
4.4 Procedure	118

Chapter	Page
4.5	Stability of response from 5TS to 4TS . . . 121
4.6	Evidence of learning over subsequent trials 128
4.7	The relative success of the two groups . . 134
4.8	Interpretation of the difference between the groups on A and Ah 136
4.9	The relative difficulty of detecting 'critical cards' when the first named character is on the reverse side 138
4.10	Ultimate success or failure in the verification task 140
4.11	Views about the truth-value of an A statement when the subject-class is empty 141
4.12	Review and discussion 142
5.	<u>The Negative Particle and Aspects of Personality</u> . . 151
5.1	Introduction 151
5.2	The negatives task: materials, procedure and subjects 152
5.3	Results 154
5.4	The negatives task, emotionality and extraversion 163
5.5	The Eysenck Personality Inventory: general aspects 166
5.6	The Eysenck Personality Inventory: administration and results 168
5.7	Summary 171
6.	<u>Review, Discussion and Prospect</u> 173
 <u>Appendices:</u>	
A	Valentine's Reasoning Tests for Higher Levels of Intelligence: a Critical Appraisal 184
B	Materials and Data Relating to the Valentine Reasoning Tests 219
C	Glossary of Logical Terms Used in this Thesis 264
D	Materials Relating to the Review of Literature . . . 266
E	Preliminary Data Relating to the GR and PR Groups . . 269
F	Materials and Raw Data Relating to the 5TS Experiment 271
G	Raw Data Relating to the 4TS Task 288
H	Raw Data Relating to the Negatives Task and E.P.I. Scores 292
 <u>Bibliography</u> 297	

SUMMARY

In an opening review of the literature it is shown that two of the factors most widely supposed to produce error in syllogistic reasoning, namely, the 'atmosphere' of the premises and the emotional significance of the content of the argument, respectively could not, and have not been shown to, have this effect. Other factors liable to produce such errors are: misunderstanding of the universal affirmative (as implying the truth of its converse), a conflict between the truth-value of a conclusion and the validity of the corresponding argument as a whole, and an inadequate grasp of, or adherence to, the 'logical task'.

In Chapter 2 the selection, on the basis of performance on two Forms of Section B of Valentine's Reasoning Tests, of a (PR) group of 22 'poor reasoners' and a (GR) group of 'good reasoners', matched one to one on a composite measure of their (superior) academic attainment and ability is described. Preliminary evidence relating to the difference between the groups is considered and the materials and procedure in the first, 'five types of statement' (5TS), experiment described. This experiment applies a form of Wason's 'card-turning task' to the three types of statement occurring most frequently in the criterion Valentine test, namely, the universal affirmative in categorical form (A), the universal negative (E) and statements of the form, 'Only X's are Y's', and to two other types incorporating logically important connectives, the universal affirmative in hypothetical form (Ah) and a universal-disjunctive (Ad).

The assumption underlying the 5TS experiment is that a difference between the groups in their grasp of any of the types of statement will be reflected in a difference in the success with which they tackle the task with respect to that statement-type. Significant differences were found on A and F and on one aspect of the response to Ah, but not on E or on Ad. A progressive analysis of responses to different statement-types suggests the operation of other, unsuspected, factors and throws doubt on the validity of the use of the card-turning task as a measure of the relative difficulty of different statement-types.

In the later (4TS) experiment, incorporating modifications to the materials and procedure, it was possible to establish (1) the persistence of the differences on A and F over a period of 12 to 18 months, (2) the superiority of the GR group in learning to make the correct response

to A and (3) the significantly greater proneness of the PR group to a type of error which can only reflect a tendency to misunderstand A statements in the way described in the first paragraph. Other aspects of the task less immediately related to the difference between the groups are also considered. Chapter 4 closes with some examples, drawn from the history of psychology, of the sometimes serious practical consequences of the misunderstanding mentioned under (3) and a discussion of possible sources of this error.

Chapter 5 describes the use of Wason's 'construction task' to investigate the possibility that PR subjects have significantly more difficulty than their GR counterparts with the negative particle, a very important element in all reasoning. The outcome is somewhat equivocal, there being a significant difference between the groups on the false-affirmative condition (which incorporates a single negative component, the 'semantic' notion of falsity) and when this condition is combined with the true-negative, but not on the false-negative which includes both negative components. This last result, it is suggested, may be due to the conspicuous instability of the response-time measure of difficulty in this case.

Chapter 5 also considers evidence (from scores on the E.P.I.) about the extraversion and emotionality of the two groups, it having been argued that differences in either of these dimensions of personality might be partly responsible for difficulties with the negative particle. Although the differences are in the expected direction none is large enough to be significant on a two-tailed test. A small but significant correlation between scores on the N scale of the E.P.I. and latency of response to the true-negative condition when both groups of subjects are taken together lends, again in the absence of corroboration from the false-negative condition, rather uncertain support to the Eifermann hypothesis of a relationship between emotionality and slowness of response to negative statements.

A concluding chapter reviews the outcomes of the above researches, considers an alternative interpretation of the 5TS results and indicates some possible points of departure for future research in this area.

CHAPTER ONE

INTRODUCTION AND CRITICAL REVIEW OF THE LITERATURE

Summary The origin of the writer's interest in individual differences in reasoning ability is briefly described. The monograph by Wilkins (1928) is reviewed both for its evidence confirming the reality of the phenomenon under investigation and also for the suggestions it contains about the possible sources of the difficulties some highly intelligent subjects seem to have with deductive reasoning tasks. A series of papers relating to the 'atmosphere effect' is then considered, the conclusion drawn being that the nature of this phenomenon has been misconstrued and that it should no longer be regarded either as non-logical or, therefore, as tending to increase the number of errors made in syllogistic reasoning tasks of the kind described. Papers attempting to relate errors in syllogistic reasoning to the subject's attitude towards the conclusion of an argument are next reviewed, it being shown (a) that there is some evidence to suggest that the subject's assumptions about the truth or falsity of the conclusion may affect his view about the validity of the argument but (b) that none of the attempts to establish a separate role for specifically emotional factors has been successful, in some cases, at least, because the researches involved are marred by serious methodological flaws. Finally, some more general discussions of the problem of error in syllogistic reasoning are considered and conclusions drawn about the most useful directions in which to conduct a search for a solution to the problem described at the beginning of the chapter.

1.1 Introduction During the years 1960 - 66 the writer held a post in the Department of Logic at the University of Glasgow. At that time every undergraduate who took the first-year course in Logic was required, as part of the work of the class, to attend a course of lectures on 'formal logic' in which the traditional Aristotelian logic was expounded and applied to examples. These lectures listed the characteristics of valid immediate inferences, syllogisms and sorites and drew the attention of members of the class to the commonest kinds of 'formal' and 'material' fallacies.¹ The members of the class were subsequently required, in exercises and examinations, to apply this training in logic to various arguments drawn from such sources as newspapers, textbooks and so on. They were to say whether an argument was valid or not and to show it to be one or the other, in the former case by setting it out in the form of an immediate inference, syllogism or sorites, with the distributions

1 A glossary of logical terms is provided in Appendix C.

of the terms marked; in the latter by naming the fallacy and, where appropriate, 'exhibiting' it by showing, for example, that the terms were not distributed in such a way as to justify drawing the conclusion from the premises.

About the usefulness of this undertaking there was a good deal of argument on points of interest to, and decidable only by, psychologists. (It was often felt, for example, that the most difficult part of the whole task was the initial translation of an argument in everyday terms into 'logical form' and that it was not uncommon for this procedure, which was supposed to make it easier to assess the validity of an argument, actually to make it more difficult - because a student who started out with a (correct) belief that an argument was valid was liable to find himself, through some mistake in translation, 'proving' it to be invalid.)

What specially interested me was the discovery, in some cases, of a large discrepancy between a student's success in this kind of task and his performance in the rest of the work of the class. In particular, some of the best students in terms of the classwork as a whole found the formal logic task surprisingly difficult. Senior colleagues had long accepted this as a fact of life and were inclined, at times, to put it down to the unsatisfactory nature of the formal logic component of the course (now in fact dropped). Other satisfactory explanations were difficult to find but I was prevented from taking the one just mentioned too seriously by the discovery, in a pilot piece of research for the Ed. B. degree, that large differences existed in rank order when 45 volunteer members of the first year Logic class were tested with Section B of Valentine's Reasoning Tests for Higher Levels of Intelligence¹ and a test of general intelligence of a more traditional kind, Heim's A. H. 5.

The phenomenon seemed the more striking because, at least on the face of it, what was involved was not simply a difference in reasoning ability as compared with overall ability or attainment but between

¹ VRT (B) was chosen for this purpose because, of all published tests, it came closest to presenting a task of the kind involved in our formal logic exercises. Because subjects are not asked, in VRT (B), to translate arguments into logical form the explanation favoured by my colleagues did not seem to apply to such discrepancies between performances on the two tests as were found.

reasoning ability as revealed by a task of the kind described above and aptitude for philosophy (the primary concern of the 'Logic' department). This made the phenomenon a particularly puzzling one because it is quite common nowadays to maintain that the merit of a philosophical training is that it teaches one to think clearly, and indeed anyone who has followed a philosophy course to the Honours stage or beyond will be inclined to agree that a philosophical training, even if it is much more, is at least a matter of learning to detect flaws in other people's arguments and to produce arguments of one's own which are without flaws.

In the research reported in this thesis, however, the focus of interest was not on this rather special kind of case of the good philosophy student whose aptitude for syllogistic reasoning is not of a comparably high order but on the more general case of the student of above average academic ability and attainment in general (as measured principally by his performance on Terman's Concept Mastery Test and on the Higher Grade of the Scottish Certificate of Education) whose performance on Section B of Valentine's Reasoning Tests (in the two forms discussed in Appendix A) was distinctly lower than one might have expected.

The reasons for the choice of the more general case were several. To begin with the most practical, it would have been extremely difficult to find a large enough pool of philosophy students in classes beyond the introductory stage from which to draw the requisite two groups of subjects of comparable, and above average, aptitude for philosophy who also possessed distinctly different aptitudes for the syllogistic reasoning task. At the same time, and more important from a theoretical point of view, the more general case, though, as I have suggested, less surprising than the special one of the good philosophy student, is still, I think, sufficiently contrary to normal expectations to merit investigation - and also, of course, of more general interest and potential practical importance.

It is not only that, I think, one naturally expects a person of high academic ability and attainment to have an equally unusual ability to detect fallacies in arguments to which he is exposed but also that there exists some evidence of a less subjective kind which seems to support this expectation. Thus the main section of Valentine's Reasoning Tests, which are offered as a test of general academic ability, consists of items of the syllogistic reasoning variety. This test may be thought to have face validity on the grounds that almost half of a sample of the

(undergraduate) population for which the Valentine test is intended regarded it as a good test of all-round intellectual ability;¹ but it must derive such theoretical support as it has² from the conclusions of, for example, Burt (1949) about the place of deductive reasoning ability in the 'structure of the mind'. Burt's conclusion is that this kind of ability, while not identifiable with intellectual aptitude in the most general sense, is yet of a very high order of generality - thus consolidating and extending a view expressed thirty years earlier (1919) that the syllogism test is the most satisfactory single test of general ability, at least for children. In the light of the later contention it seems fair to see as a phenomenon calling for further investigation those cases in which a subject's performance on a measure of deductive reasoning ability is decidedly poorer than one would expect on the basis of his overall attainment and aptitude as measured by other tests of 'intelligence' and by performance on more orthodox tests of scholastic aptitude.

1.2 Some prefatory remarks on the relevant literature Apart from the paper by Wilkins (1928) to be discussed in the next section I have been unable to find any references in published work to the problem dealt with in this thesis, viz., the sources of individual differences in deductive reasoning ability in adult subjects of high intelligence. On the other hand, a rather limited amount of research has been done into the sources of error in syllogistic reasoning in the undergraduate population at large. To the extent that such sources have been identified with some certainty it is possible in principle to generate hypotheses as to the factors responsible for individual differences on the assumption that relatively poor reasoners will be more susceptible to the sources identified than relatively good reasoners. Thus, for example, if it were established that an 'atmosphere effect' is generally liable to make fallacious arguments appear to be valid, then one might investigate the possibility that relatively poor reasoners are especially subject to such an effect. It was with this possibility in mind that I looked to the other papers reviewed in this chapter to supplement the rather casual suggestions contained in the paper by Wilkins.

1 See Section A.7 in Appendix A.

2 In Section A.8 of Appendix A I shall consider the theoretical basis on which the Valentine test must be supposed to rest and suggest that, in the final analysis, it is not a satisfactory one. While this conclusion depends on the view that the relationship between syllogistic reasoning ability and general ability is not a sufficiently close one, I think that the interest of the question why it is not a closer one remains.

1.3 Wilkins on the effect of changing the material in terms of which a reasoning task is presented As its title implies, the long paper by Wilkins (1928) focuses on a problem quite different from the one with which this part of this thesis is concerned. At the same time Wilkins devotes a quite substantial amount of space to the consideration of the relationship between syllogistic reasoning ability and general academic aptitude as measured by Thorndike's College Entrance Examination. There are, as a result, at least three reasons for placing a discussion of this paper at the beginning of the present review of literature. It is, to begin with, the earliest discussion in English of the problem of error in syllogistic reasoning¹. It reports work arising from observations strikingly similar to those which prompted the research described in this thesis. And, partly as a direct outcome of these observations, partly as a consequence of Wilkins's investigation of the relationship between syllogistic reasoning ability and general academic aptitude, it contains the only published discussion I have been able to find of the striking kind of individual differences in syllogistic reasoning ability with which this part of this thesis is concerned.

On page 5 of her paper Wilkins remarks:

A college class in logic seems to divide itself very soon into two groups, those who seize on the material given at once and in a few hours understand everything that it takes weeks for the other group to assimilate, if it ever does assimilate them. This second group seems to understand principles fairly well as long as they are applied to facts within their experience, but seems unable to apply any principles as soon as the material is unfamiliar or symbolic. For instance, they see clearly that if you have the proposition, 'All horses are animals' you cannot logically deduce from that the proposition that, 'All animals are horses'; but whenever the proposition is, 'All x's are y's' they feel sure that this necessarily implies, 'All y's are x's'. Many of the members of this second group stand above the median in most of their classes. Some of them have unusual verbal facility.

This passage is of interest not only because it shows that Wilkins was impressed by a failure on the part of some of her brighter students altogether comparable with the one in which my colleagues and

1 Sells (1935) refers to articles by Eidens and the two Störrings in Arch. f. d. ges. Psych. from 1908 onwards. In fact a rather superficial examination of this journal over the period in question revealed a considerable number of papers by a variety of authors containing material, mainly of an introspective kind, which appeared to be of relevance to the issues discussed in this thesis. Unfortunately, the present writer's German prevented a more thorough scrutiny of these papers.

I were interested but also because it touches on two factors which appear to play an important role in the production of errors in syllogistic reasoning, one of them being, as I hope to show later, at least partly responsible for the difference between the relatively poor reasoner and the relatively good one. The two factors are the reasoner's beliefs about the truth or falsity of the conclusion and his understanding - or misunderstanding - of the universal affirmative form of statement. As to the existence of a group of undergraduates of superior all round ability who make an unexpectedly poor showing on a syllogistic reasoning task, Wilkins is able to provide more than simply her own impressions in support of my own. Taking her subjects as a whole she found correlations of about .50 between the various Parts of her Syllogism Test and the College Entrance Examination. On her own submission this degree of relationship must be regarded as reasonably high in view of the homogeneity of her sample. At the same time there were some subjects whose score on the Syllogism Test was surprisingly high considering their scores on the C.E.E.

On the other hand, much more conspicuously out of line with the general trend of the correlation table is a number of cases that seem to form almost a separate group. These are mostly just above the median of the intelligence scores and very low in the syllogism test. These seem to be individuals of good general intelligence who are able to succeed but poorly in formal syllogistic reasoning. One individual who on the intelligence test has only two individuals making a better score, on the syllogism test has thirty-nine individuals [out of sixty-nine] making a better score. This group stands out conspicuously on all parts of the test except Part B [where] the scores of these individuals were not actually higher than in the other parts, but only relatively so.

ibid. p. 30

Reference is made in this passage to various 'Parts' of the Syllogism Test used by Wilkins. Since these relate not only to the issue referred to in the title of her paper - the effect of changing the materials in terms of which a reasoning task is presented - but also to the first of the two sources of error mentioned in the previous paragraph, it now seems appropriate to say something more about them.

In the passage from page 5 quoted above Wilkins says that a certain category of undergraduate gets into difficulties in a logical reasoning task if, in an argument of a given form, 'familiar' material is replaced by material which is 'symbolic' or 'unfamiliar'. Other, more objective evidence to this effect had previously been advanced by Thorndike (1922) using algebraic reasoning tasks. Wilkins's Syllogism

Test was designed in such a way as to provide additional evidence on this point but with specific reference to syllogistic reasoning.

There were four parts to the test, the syllogisms and immediate inferences in each being identical in 'form' while differing in 'content' - in the materials of which they were composed. In Part A the terms of the syllogisms were familiar words and the conclusions were framed in such a way as to be neither obviously true or false. (For example, all Mary's cats, rather than simply - and falsely - all cats, were said in the conclusion of one syllogism to be black.) In Part B the terms of the syllogisms were letters, as, for example, in 'all X's are Y's'. Wilkins refers to these as syllogisms in abstract form. In Part C the syllogisms' terms were unfamiliar (and frequently rather complicated) words such as 'tuscambia' and 'tiksatopses'. Finally, in Part D the syllogisms were composed of what Wilkins calls 'suggestive' materials. For the most part this means that they were such that the truth-value of their conclusions was at odds with the validity of the argument as a whole: a valid argument would have a conclusion which was unmistakably false, an invalid argument a conclusion which was unmistakably true.

On the whole, Wilkins found, as expected, that the syllogisms of Part A were easier than those of any other Part. That is to say, she seems to have shown that syllogisms are harder to judge if they are presented in abstract terms, in unfamiliar and complicated words, or if the truth-value of their conclusions is at variance with the validity of the syllogisms as a whole.¹

As we have seen, in the quotation from page 30, Wilkins found a conspicuous tendency for one group of subjects to do much less well than their scores on the College Entrance Examination would have led one to expect. This was true for all Parts of the test except B where they seem to have acquired a kind of anonymity by virtue of the fact that all subjects did less well on this Part than on the others and so were bunched together more. In the circumstances these results seem

¹ In the quotation from page 5 Wilkins seems to be confusing the second and third of these effects. The conclusion, 'All animals are horses' is obviously false (which helps the subject to recognise that it doesn't follow from 'All horses are animals'); the terms employed, on the other hand, are entirely familiar ones.

to suggest - in accordance with the general principle stated in 1.2 above - the generation of two hypotheses as to the sources of individual differences in syllogistic reasoning ability: that poor reasoners are relatively more affected than good reasoners by the presentation of arguments in unfamiliar and complex terms, and that they are more prone to be affected by a combination of a valid argument with a false conclusion or an invalid argument with a true conclusion.

Since the differences in syllogistic reasoning ability with which the research reported in this thesis is concerned relate to arguments of the kind included in Section B of Valentine's Reasoning Tests, however, the first of these hypotheses may, I think, be dismissed without further ado, for the arguments of that section are not in unfamiliar or in complex terms. The second possibility calls for closer consideration, not only because it is not so obviously irrelevant to our criterion test but also because, as we shall see in section 1.5, it refers to an effect which has been established by other researchers besides Wilkins. I shall defer a full discussion of the extent to which it does bear on performance in the Valentine test until that section, in the meantime simply recording^{my belief} that it is unlikely to account for more than a relatively small part of the individual differences in deductive reasoning ability with which this thesis concerns itself.

Wilkins found that of the two types of task involved in her Syllogism Test - the detection of fallacies and the recognition of valid inferences as valid - the former represented the better means of discriminating between subjects. It was possible, moreover, to show that certain types of fallacy are harder to detect than others and that some (not necessarily the same ones) discriminate better than others. In the Valentine test, Section B an important part of the discriminatory power depends on a third type of task not represented in Wilkins's test, the recognition of a statement of the flaw in an invalid argument. To the extent that the Valentine test does depend for its effectiveness on the subject's need to detect fallacies, however, it is clearly relevant to the task of this thesis to consider any factors which interfere with a subject's ability to carry out this task successfully, and especially in the case of fallacies with superior discriminating power.

It is in this connexion that we now return to the first of the

the possible sources of error mentioned by Wilkins in the first of the two quotations given above - the tendency for subjects to suppose that a statement of the form, 'all x's are y's' means or implies that all y's are x's. Wilkins's reference to it in this passage is, of course, a rather casual one, but she returns to it on page 72 where she offers it as an example of the 'simplest and commonest fallacies of thought' of which 'many college students' are entirely unaware. Our interest in this fallacy is not simply for its own sake but also for the role which it can be thought to play in other fallacies, including some of those which proved, in Wilkins's test, to be most difficult to detect and most effective in discriminating poor reasoners from good ones.

The fallacies in question are those of the Undistributed Middle Term, the Illicit Process of the Major Term, and the Illicit Process of the Minor Term, ranking respectively second, third and fifth in difficulty and first, fifth and fourth in discriminating power. Of the large number of syllogisms incorporating these fallacies eighteen of the most persuasive can be represented summarily as follows (using the traditional symbols for subject, predicate and middle terms):-

A. Undistributed Middle

Some M is (not) P	All (some) P is M	All P is M
All S is M	All (some) S is M	Some M is (not) S
∴ Some S is (not) P	∴ All (some) S is P	∴ Some S is (not) P

B. Illicit Minor

All (no) M is P	All (no) P is M
All M is S	All M is S
∴ All (no) S is P	∴ All (no) S is P

C. Illicit Major

All (some) M is P	All (some) M is P
No (some) S is (not) M	No (some) M is (not) S
∴ No (some) S is (not) P	∴ No (some) S is (not) P ¹

The words in brackets in the above tabulation indicate alternatives and are self-explanatory except in the respect that of course 'some' can be substituted for 'all' and 'some...not' for 'no' in only one premiss at a time if the syllogism is not to contain another fallacy in addition to the one in question, that of drawing a conclusion from two particular premises. Inspection of the above tabulation will

¹ I owe this tabulation of the plausible forms of these fallacies to my former colleague, Mr Eric Toms.

reveal that of the eighteen different forms of the syllogism there represented all but two are liable to go undetected by a subject who assumes that 'all S is P' implies that all P is S.

In the case of the fallacy of the Undistributed Middle Term what is wrong with the syllogism is that the premises do not assert anything about the whole class of things represented by the M term. The result is that one has no right to draw any conclusion about the relationship existing between the classes represented by the S and P terms: all horses and all pigs may be animals without any pigs being horses or, of course, vice versa. However, if, as Wilkins suggests, some subjects suppose that 'all S is M' implies that all M is S, then they would naturally read a syllogism of the above kind as asserting something about the whole class represented by M: if all horses are animals, then, it is assumed by a person guilty of the above confusion, all animals are horses and this, together with the statement that all pigs are animals does imply that all pigs are horses - except that, as Wilkins points out, the obvious falsity of the converse of 'all horses are animals' is likely to prevent the occurrence of the error in this particular case.

Similar considerations apply in the case of the Illicit Process of the Minor and Major Terms. In all four of the examples given, the fallacy disappears in the Minor case if the universal affirmative is assumed to mean or imply not only that all x's are y's but also that all y's are x's. The forms cited are fallacies because the conclusion asserts something about the whole class represented by the S term when nothing is said about this whole class in the minor premiss. Given the tendency to interpret a universal affirmative as asserting a symmetrical relationship between the two classes mentioned in the premises, however, the inference is readily mistaken for a valid one.

Two of the forms that the fallacy of Illicit Process of the Major can take are exceptions to the rule that all the fallacies represented in the above list would escape detection by a subject who thought that 'all x's are y's' implies its converse. These are the ones in which the major premiss is particular affirmative and the minor premiss universal negative. In the other four cases the conclusion asserts something about the whole class represented by P although nothing is asserted about this whole class in the major premiss - at least as

long as one does not make the illicit conversion we have been talking about.

So much for the potential importance of this particular error. Much of this is recognised by Chapman and Chapman (1959) who include eleven of the sixteen cases mentioned above in their list of errors in syllogistic reasoning which can be explained on the assumption that subjects illicitly convert A and O propositions. Two questions remain: what evidence there is that such a confusion is common amongst undergraduates, and how large a role it is likely to play in the individual differences in performance on Section B of the Valentine Reasoning Tests with which this part of this thesis concerns itself.

As to the second of these questions, the answer is that three¹ of the six fallacious arguments in each form of this Section would not appear to be fallacious to subjects supposing that the truth of 'All x's are y's' guarantees the truth of the statement that all y's are x's. The same would hold of a fourth argument (item 14) if we adopted the traditional logician's view that a statement of the form 'Only x's are y's' is equivalent to a statement of the form 'All y's are x's'. In fact, while not, of course, disputing the logical equivalence of these two forms of statement, I doubt their psychological equivalence and will be presenting evidence in a later chapter in support of this view.

As to the incidence of the confusion in the undergraduate population, we have, of course, Wilkins's own finding that the fallacy of illicit conversion of an A proposition frequently goes undetected, especially when combined with obversion. Woodworth and Sells (1935), too, take the frequency of this kind of error for granted, partly in the light of Wilkins's discoveries and partly following Eidens. Indeed the 'atmosphere effect', which we shall be considering in some detail in the next section, was apparently originally introduced by Woodworth to account for the tendency to make illicit conversions of A and O propositions. Chapman and Chapman try, in my view rather unconvincingly, to show that this tendency is only to be expected in view of the fact that an A or O proposition and its converse are usually as a matter of experience either both true or both false.²

1 Items 5, 7 and 8 in the original form, 5, 6 and 10 in 'Form W'.

2 Chapman and Chapman attribute the failure to detect two further fallacies to the tendency to make the illicit O conversion.

Wilkins's study, then, provides the only direct evidence I have been able to discover that undergraduates do make the kind of confusion about the significance of statements of the form, 'All x's are y's' which we have seen to be potentially important for errors in syllogistic reasoning. Even her evidence is not without its problematic aspects, however. She reports that 26 per cent of her subjects said they thought it possible to infer that all those making high grades in the intelligence tests are good students given that all good students make high grades in the intelligence tests. However, the incidence of the error in this case is likely to be affected by two factors - operating in opposite directions: in the first place, the incidence may have been reduced by the fact that subjects would be on their guards against such a fallacy, and in the second place, it may have been increased by the plausibility of the conclusion. In a later chapter I shall present evidence of a different kind to which neither of these objections seem to apply.

I have already referred to possible confusions about the significance of two types of statement commonly figuring in the traditional logic. Wilkins refers to a third, the tendency to assume that 'Some x's are y's' implies that some x's are not y's and, conversely, that 'Some x's are not y's' implies that some are. The importance of this particular confusion is, not that it makes subjects miss a fallacy present in an argument, but that it leads them to suppose a fallacy present in an argument which is logically impeccable. In its sharpest form this would lead a person to deny that 'Some x's are y's' is consistent with 'All x's are y's' and to deny that 'Some x's are not y's' follows from 'No x's are y's'. As Wilkins remarks, this rather surprising kind of confusion is probably to be explained in terms of the fact that we do sometimes use 'some' in such a way as to indicate that we know or suspect that what we say about some members of the class in question may not said about all members of the class. Where we seek to make our meaning on this point unambiguous, we may resort to the use of expressions such as 'at least some' or 'some if not all' on the one hand and 'only some' on the other. In the light of this finding of Wilkins subsequent studies involving syllogistic reasoning include a careful clarification of the point. There is no such explanation in the Valentine Reasoning Tests, but it seems possible to disregard this as a source of error in that test because the word 'some' is used only once - and then with the saving words, 'at least'. A complete list of the types of statement used in the Valentine test is given in Appendix B.

1.4 The so-called 'atmosphere effect' After Wilkins the next psychologists to discuss factors related to error in syllogistic reasoning were Woodworth and Sells (1935). They mentioned three such factors including, as the first of these, the confusion about the meaning of 'some' just discussed. The second proposed factor was 'caution', a tendency, in an experimental situation, to 'play safe' by preferring a particular conclusion to a universal one and, more problematically, a negative to an affirmative one. Their paper is chiefly remembered, however, for the third factor they proposed, the effect of what they called the 'atmosphere' of the premisses.

Working within the traditional fourfold classification of the syllogism as universal affirmative (A), universal negative (E), particular affirmative (I) and particular negative (O), they suggested that the atmosphere of each type of premiss is as follows:-

A statements have an all-yes atmosphere
 E statements have an all-no atmosphere
 I statements have a some-yes atmosphere
 O statements have a some-no atmosphere

According to Woodworth and Sells (1935) this notion of an atmosphere was introduced by Woodworth originally to account for the fact that subjects tend to assume that:-

'All x's are y's' implies 'All y's are x's'
 'No x's are y's' implies 'No y's are x's'
 'Some x's are y's' implies 'Some y's are x's'
 'Some x's are not y's' implies 'Some y's are not x's'

Subjects who make this assumption are said to do so because the members of each pair are more like one another, in quantity and quality, than they are like any other statement - in other words, because their atmospheres are as similar as possible.¹

We have already had occasion to discuss the first and fourth of the above inferences, and, as they are illicit conversions, it is no doubt necessary to try to explain why undergraduates are apparently prone to make them. No such need appears to arise in the other two cases, however, for these are justified by logic. In any case, if the atmosphere effect explained only these phenomena, it would scarcely deserve the attention it has received, for there seems to be no reason to prefer the atmosphere hypothesis to an appeal to one other non-logical aspect of the pairs given above, namely their symmetry or reciprocity.

¹ For all this see Woodworth and Sells (1935) pp. 452-3 and Sells (1936) pp. 12-13.

In practice, however, the atmosphere effect achieved prominence because it seemed to make it possible to predict patterns of error not only when premises of the same atmosphere but when premises of different atmospheres were combined to produce a syllogism as opposed to an immediate inference. To this end certain 'secondary hypotheses' had to be introduced. These were (Woodworth and Sells, 1935, p. 454):

(1) that a particular premise creates a some atmosphere, even though the other premise be universal, and (2) that a negative premise creates a negative atmosphere even though the other premise be affirmative.... In detail, the secondary hypotheses are that:

With premises AA atmosphere calls for an A conclusion

With premises AE or EE atmosphere calls for an E conclusion

With premises AI or II it calls for an A conclusion

With premises AO, EI, EO, IO, or OO it calls for an O conclusion

Sells (1936) tries to defend these secondary hypotheses. He had begun his paper by developing a suggestion made in the earlier paper with Woodworth that 'atmosphere' is a fairly, or rather very, extensive phenomenon, similar, in general character, to 'set' or 'Einstellung'. As such 'the result of the atmosphere effect is that the individual makes a response (e.g. an inference or judgement) which is most similar in quality to the general trend or tone of the whole situation set up.' (p. 8) Proposed examples of the atmosphere effect in this broader sense include experimental or quasi-experimental phenomena, such as the 'halo' effect in rating, Thorndike's 'spread of effect', and so on, the common feature of which is, as we have been led to expect, simply that some kind of perceived similarity between different stimulus situations seems to be important in determining the subject's responses, not a point of great novelty even in 1936.

So far as the secondary hypotheses relating to syllogistic reasoning are concerned, Sells says (p. 15):

If one premise is A or E and the other I or O, the atmosphere is partly universal and partly particular, a blend of all and some, which would certainly be weaker than a straight all and would thus amount to a some; and if one premise is A or I and the other E or O, the atmosphere is partly affirmative and partly negative, which would be weaker than a straight yes and would thus amount to a no.

The defence of the secondary hypotheses offered here is a very unconvincing one, particularly so far as the second of them is concerned. Obligated to choose simply between 'some' and 'all' (as of course one is, within the framework of the traditional syllogism) one might very well choose the former as the best blend of the two. For although it hardly seems a satisfactory compromise between two alternatives to opt for one of them,

at least 'some' can be used to refer to nearly all of a class as well as to hardly any. On the other hand, the choice of the negative rather than the affirmative as a compromise or blend between them seems entirely without justification: it seems perfectly reasonable to turn Sells's argument upside down and say that, as a yes plus a no is stronger than a no by itself, it amounts to a yes.

In fact, Sells appears to be aware of the dubious nature of the basis here proposed for the two hypotheses. He says (p. 16)

The atmosphere effect is defined as a set to complete a task with that one of several alternative responses which is most similar to the general trend or tone of the whole problem. It seems arbitrary, then, to say the least, to resolve the affirmative-negative and particular-universal cases by stating that the tone of the whole problem in the former case is negative and in the latter case particular. However, this procedure is neither arbitrary nor 'a posteriori'. It rests upon the same foundation as the rules for the validity of the syllogism which state:

- a) if one premise be negative, the conclusion must be negative and
- b) if one premise be particular, the conclusion must be particular.

This seems to me to be a major concession because it substitutes an essentially logical basis for the atmosphere effect for the non-logical one of similarity, and it is a major concession because it seems most unlikely that the atmosphere effect would have received the attention it has if it had not been understood to represent the operation, in syllogistic reasoning, of some aspect of the situation - namely 'similarity' between premises and conclusion - totally irrelevant to the task of logical appraisal.¹

As it is, Sells continues to talk, in his discussion of Table 1.1 below, as if the degree of similarity between premises and conclusion were the factor operating to determine the subjects' choices. Commenting on the fourth and fifth rows, for example, which show that, presented with a pair of affirmative premises, one particular and one universal, Sells's subjects were most likely to choose an I conclusion, followed by an A and O conclusion (roughly equally), followed by the E conclusion, Sells remarks (p.34), 'Both A and O propositions are more similar to I than the E proposition, for, since I is particular affirmative, A resembles I in being affirmative, O resembles I in being particular, while E is neither particular nor affirmative'.

¹ For evidence about the interpretation put upon the atmosphere effect by later writers see p. 17 below.

TABLE 1.1

The Effect of Atmosphere on the Acceptance of Invalid Conclusions

(From Sells (1936), p. 35) Entries indicate the average percent accepting each type of false conclusion. N = 65

Premises	Types of Invalid Conclusion			
	A	E	I	O
AA	58	14	63	17
AE	11	51	13	63
EA	8	64	12	69
AI	33	4	70	32
IA	36	15	75	36
AO	15	26	42	76
OA	13	33	28	75
EE	21	38	25	34
EI	8	40	22	62
IE	11	42	22	63
EO	13	29	29	44
OE	15	31	24	48
II	27	9	72	38
IO	12	19	31	64
OI	11	23	33	71
OO	14	16	38	52

In the application of the similarity principle in this case, however, Sells's procedure is arbitrary in a new sense, for he simply ignores one of the premises, namely the A premise. If it is taken into account, a proper estimate of the 'similarity' of the four alternative conclusions to the premises might be set out as follows:

<u>Conclusions</u>	<u>Premises</u>			
	A		I	
	Universal	Affirm.	Partic.	Affirm.
A	+	+	-	+
E	+	-	-	-
I	-	+	+	+
O	-	-	+	-

Such an estimate shows that A and I agree with the premises on three points out of four, E and O on only one. Hence we should expect, on the basis of similarity alone, that A and I are equally, and more often, preferred, E and O equally and less often. Estimates of this kind of the degree of similarity between other pairs of premises and the four possible conclusions reveal a comparable discrepancy between the outcomes predicted on the basis of similarity and the outcomes actually reported in the above table, and predicted on the essentially logical 'secondary

hypotheses'. Thus, for example, in the case of syllogisms with an A and an O, or with an E and an I, premiss, the distribution of choices should approximate to a chance distribution since all four types of conclusion resemble the premises on two counts only. In fact, there is a pronounced tendency for the O conclusion to be preferred in both cases.

It will be obvious that the effect of atmosphere need not be contra-logical even if it were non-logical - as it would be if it were based on a perceived similarity between premises and preferred conclusion. On the other hand the assumption that it is contra-logical has been made from the very first. Sells himself says (page 40): 'The atmosphere effect is a directive tendency which is opposed to the perception of the correct logical relations in these invalid problems.' The natural consequence of such a factor is, of course, error in syllogistic reasoning, and there is ample evidence that atmosphere has come to be regarded as the source of such error. (Cp Janis and Frick (1943), Morgan and Morton (1944), and, most recently, Vinacke (1957) who, commenting on Sells's results, remarks (p. 83): 'These results are very important for everyday reasoning where deduction is concerned, since they reveal a source of error apart from the facts presented in the premises.')

As I understand them, however, Woodworth and Sells thought of the atmosphere effect as a factor explaining why one kind of error is more likely to be made than any other and not why an erroneous conclusion is preferred to a correct one. It is true that atmosphere was held to explain why a subject is likely to say that an A conclusion follows from two A premises and deny that the corresponding E conclusion follows - even though both are actually invalid. In this sense the atmosphere effect may appear to be a source of error, but the error is a 'good' one because a subject who accepts an A conclusion and rejects an E one obviously recognises that any conclusion which follows from two affirmative premises must itself be affirmative - one of the rules of the syllogism. Similarly, subjects whose choices reflect the effects described in the 'secondary hypotheses' must be assumed to recognise - at some level¹ - that the conclusion must be negative if one premise is negative and particular if one premise is particular. Their failures are due to the fact that they do not realise that there are other rules of the syllogism - notably, of course, the rules of distribution.

¹ Unconsciously, as one would expect and as the results of Sells's introspective study (pp. 47 ff.) tend to confirm.

If the atmosphere effect reflects an awareness of some of the conditions which have to be met if a syllogism is to be valid, one ought perhaps to expect that the effect will be more obvious in the case of good reasoners than in the case of poor ones. An outcome opposite to the one implied by the assumption that the effect of atmosphere is contra-logical, it is clearly one which I could have tested empirically with my two groups of subjects. At this point I have to admit, however, that while I was planning my experimental work, I was still under the impression that atmosphere accounts for kinds of error in syllogistic^{reasoning} and is thus unrelated to such individual differences in syllogistic reasoning ability as are revealed by the Valentine test. It came as a very pleasant surprise, therefore, to discover that there is in Sells's own results some tentative support for this conclusion.

Sells was interested in the extent to which the atmosphere effect is related to age and intelligence. It isⁱⁿ connexion with the second of these two variables that the results germane to the present question are presented. Sells compared the susceptibility to the atmosphere effect of two groups of widely different intelligence as measured by the CAVD, groups which he called 'bright' and 'dull' (N=16 in each). For present purposes the interesting thing is that the two groups differed also in their syllogistic reasoning ability as indicated by the number of errors made in Sells's test.¹ The result was that the bright group appeared to be more susceptible to the atmosphere effect than the dull group.

Given Sells's view of this effect as contra-logical this outcome naturally seemed to him something of a paradox (cp. p. 40). He attempted to resolve the paradox by suggesting that the strength of the effect in the case of the dull group was 'blanketed' by another effect - what he called 'gullibility' or the tendency to suppose that any conclusion follows from any pair of premises. In my view, the proper explanation of this result is that the dull group, unaware of the rules of the syllogism represented by the atmosphere effect, were liable to regard any conclusion as equally likely to follow from a pair of premises as any other. Providing, of course, that subjects unaware of these rules are also unaware of any of the other rules of the syllogism, we might expect the pattern of their responses to approximate to a random one. In fact, inspection of the table of results in Chapman and Chapman (1959)²

1 ^{average} The/number of errors being, respectively, 27 and 74 (out of 180).
 2 Page 223, reproduced in Appendix D below.

suggests that there are other rules of the syllogism which are more obvious than those represented by the atmosphere effect.

In this table the Chapmans report on the percentage of cases in which subjects chose the four types of statement as following validly from pairs of premises from which in fact no conclusion can validly be inferred, as well as the percentage of cases in which subjects chose the correct alternative 'None of these'. A striking aspect of this table is the way in which this latter percentage increases in those cases where both premises are negative or both are particular - or, of course, both. This cannot, I think, be supposed to mean that it is easier to perceive the relationships between classes when these are expressed in negative terms - as they predominantly are in this second half of the Chapmans' table - in view of Wason's (1959 and 1961) evidence about the special difficulties experienced by undergraduate subjects with the processing of information stated in negative terms. Much more plausible, in my view, is the assumption that subjects recognised that the pairs of premises in this second half of the test violated the rules of the syllogism which state that at least one premise must be affirmative and at least one universal if any conclusion is to follow from them.

The rules of the syllogism may be supposed to form a hierarchy of obviousness, extending from the two rules just mentioned at one end, through the rules reflected in the atmosphere effect to the rules of distribution. Asked to say whether a conclusion follows from a pair of premises, a subject may be supposed to respond with an affirmative if the syllogism does not contravene any of the rules with which he is familiar. This means, amongst other things, that a subject operating with an incomplete set of rules will score some lucky hits - as well, of course, as making some inevitable mistakes. A syllogism in the AII mood may be accepted in the first figure (where it is valid) as well as in the second (where it is not) by a subject whose responses are guided only by the rules which are more obvious than those relating to distribution. In the first case he may be said, with Woodworth and Sells (p. 458), to have been lucky. On the other hand his success is not pure luck for he will reject syllogisms in the moods AIA and AIO (each of which break one of the rules of which he is aware) and, of course, AIE (which breaks two). Perhaps it is only necessary to repeat at this point that I do not suppose that subjects who have not had a course, or done some reading, in formal logic would be able, necessarily, to formulate the rules in

question explicitly. The assumption is, rather, that over the long period of learning which precedes adulthood subjects acquire what looks like a kind of intuitive understanding of the conditions which must be met if an argument is to be deductively valid. In view of the individual differences which make the subject-matter of the major part of this thesis, it may seem that some people learn this kind of thing better than others, even when allowance is made for differences in general ability and attainment. To this we shall have to return later.

I have referred on a number of occasions, in this section and in the previous one to the paper by Chapman and Chapman (1959) in which, as the title tells us, the atmosphere effect is reconsidered. It seems appropriate, before we leave this discussion of the atmosphere effect, to refer to the central contention of the Chapmans' paper, which is that the atmosphere hypothesis explains the pattern of errors in a syllogistic reasoning task rather less well than its authors supposed. Instead, the Chapmans propose two other hypotheses which together, the authors maintain, explain this pattern better. With one of these hypotheses we are familiar from the previous section, for it is that subjects are frequently unaware of the invalidity of inferring 'All y's are x's' from 'All x's are y's' and 'Some y's are not x's' from 'Some x's are not y's'. We have seen how rich a source of error the first, at least, of these misapprehensions is likely to be, and also some of the evidence that it is in fact a common misapprehension.

About the other hypothesis there must be a good deal less enthusiasm for it seems to be inadequately made out and inherently implausible. This second hypothesis is that subjects may have used 'probabilistic', as opposed to strictly deductive, inference since they 'had no way of knowing that all but strict deductive reasoning is disallowed in the syllogistic game' (p. 224). To this it is difficult not to protest that the Chapmans' subjects ought to have had a way of knowing that the task they were being asked to perform was not one of assessing probabilities but of saying which conclusion must be true if a particular pair of premises is true. No detailed information is provided about the instructions given to subjects, but there is at least no good evidence to suggest that it is impossible to make the nature of the task in the 'syllogism game' clear. (See, for example, Wilkins (1928) p. 15 and Henle (1956) p. 125.)

There are other, more important, reasons for doubting the adequacy of this second of the Chapmans' hypotheses. It is not simply that the example of a 'not unreasonable' but deductively invalid inference in the sphere of science is very unconvincing¹ but also that the alleged predictive power of this second hypothesis is not made out. As the Chapmans state it, in the kind of probabilistic inference in question "S reasons that things that have common qualities or effects are likely to be the same kinds of things, but things that lack common qualities or effects are not likely to be the same. In the syllogism, the available common characteristic is the middle term." (p. 224) It will be apparent that the Chapmans at least do not mention, even if they do recognise, that the statements of a syllogism refer, not to things, but to a part or to the whole of a class of things. This means that the probability that the things mentioned in the premises are the same kind of thing will vary from the case where both premises are universal and so refer to every thing of the kinds mentioned, through the case where one premise is universal to the case where neither is. Also the paradigm described above applies straightforwardly only to those syllogistic figures in which the middle term is predicate in both premises or may be made predicate by conversion.

Applying the hypothesis to a particular case, the Chapmans argue as follows (p. 225): "In the case of an I coupled with an O premise in the second figure, for example, 'Some A's are B's, some C's are not B's', S reasons that some A's and some C's do not share the common quality of B and therefore some C's are not A's Probabilistic reasoning yields analogous results for the case of an I with an E." This last remark does not appear to be true for, as far as I can make out, a person reasoning along 'probabilistic' lines of the kind indicated would conclude from the premises that 'Some A's are C's and no B's are C's' that some B's are not A's and not that none are - whereas, in the the Chapmans' results, it is the latter conclusion and not the former which is preferred. (The percentages are 62, 59 and 48 for E in the

1 "A chemist might reason as follows: 'Yellow and powdery material has often in the past been sulphur. Some of these test tubes have yellow and powdery material. Therefore some of these test tubes contain sulphur.'" (p. 22) Note that the first premise differs from a typical syllogistic premise in the respect that it says that yellow, powdery material has often been sulphur and not merely that it sometimes has. Even so, the unqualified conclusion that some of the test tubes contain sulphur (not may do so) is one which no self-respecting scientist is likely to draw.

relevant three items compared with 13, 16 and 24 for 0.)

Besides, although it is difficult to be certain about this, it seems that the probabilistic inference sketched by the Chapmans should favour another, unobserved result. For if the possession of common characteristics makes it likely that two classes of things have some members in common, equally, one might suppose should the non-possession of such characteristics. Thus, from the premises 'No A's are C's and some B's are not C's', one ought to conclude, on probabilistic grounds, that some B's are A's. The trouble with this kind of discussion is, of course, that a difference of opinion about what is likely to be inferred from what is capable of being settled neither by an appeal to rules of inference (as in deductive reasoning) nor, as far as I can see, by an appeal to the facts. I propose to leave Chapman and Chapman's paper, therefore, simply with the remark that they have not persuaded me, at least, that they have been able to provide a better basis for the prediction of errors in the kind of syllogistic reasoning task which they describe than the atmosphere principles set out by Sells and Woodworth.

In concluding this section it would be appropriate to consider the question whether there is any point in continuing to speak of an 'atmosphere effect' in reasoning if, as I have suggested, the nature of this effect has been misconstrued even by its originators in such a way as to mislead anyone who takes the term they have chosen to describe this effect seriously. Relatively recently Hunter (1957a, 1957b) has spoken of an atmosphere effect in reasoning tasks of a non-syllogistic character. By this he means to refer to three different kinds of case in each of which the subjects' responses seem to be determined by the 'general feel' or 'global impression' of the problem. A similar interpretation of the atmosphere effect is adopted, with reference to syllogistic reasoning, by Gorden (1953).

The weakness of such an interpretation is that it advances our understanding of the sources of failure in a reasoning task only very little. It does not seem possible to identify the 'general feel' or 'global impression' of a problem independently and so establish the alleged relationship empirically. At most the invocation of an atmosphere effect in such terms seems to mean that subjects fail to solve their problems because they fail to carry out an adequate analysis of the terms of the problems - though even this, as a falsifiable proposition, seems open

to doubt in view of the large proportion of Hunter's undergraduate subjects who said that they had 'reasoned out' their incorrect solutions.

As we have seen, although Woodworth and Sells do talk about the atmosphere effect, so far as this is apparent in processes other than syllogistic reasoning, as if it depended on some kind of 'global' or general impression which is to be contrasted with a clear and adequately analysed perception of the important relationships involved, they relate it explicitly to the degree^{of similarity} which exists between premises and preferred conclusion in the syllogistic case. (As I have already remarked, similarity between stimulus situation and response also seems to be the factor common to the various phenomena listed by Sells as exemplifying the atmosphere^{effect}/outside the sphere of reasoning.) In cases where the premises of a syllogism differ in quantity or quality I have suggested that this appeal to maximum similarity is misleading, but in other cases, where the conclusion has the same quantity and quality as both premises, the existence of a very high degree of similarity is indisputable even if its causal effectiveness remains problematic. As it happens, very much the same could be said about the problem which Hunter set for his undergraduates (1957a): the preferred (and incorrect) solution is like both premises in the respect that they all refer to a quantity which is supposed to be held constant, and to two other quantities one of which varies in direct proportion to the other.

In the other kind of problem described by Hunter (1957b) - a three term series problem administered to 11 and 16 year olds - no comparable similarity between premises and conclusion appears to exist. Instead, Hunter suggests that any atmosphere effect which appears to be operative must be ascribed to two different sources, to what Burt (1919) has called 'direct statement' or to 'inclusiveness'. It is unnecessary for present purposes to explain these in detail. The first seems to depend on a kind of perseveration: a section of the given 'rings in the child's memory' and facilitates the emission of the appropriate response. Inclusiveness involves the mistaken assumption that a statement such as 'A is darker than B' implies that A and B are dark, together with the further assumption that 'A is dark and fair' (which is a conclusion drawn by the subject from the premises) means that a is medium dark.

In fact Hunter found little evidence to support the view that 'atmosphere' from either of these sources plays an important role in

determining subjects' responses. My point would be that even if he had, it would hardly have helped to describe the result in terms of 'atmosphere' when it can be explained more precisely and unambiguously in terms of such processes as are represented by the terms 'direct statement' or 'inclusiveness'. With the availability of explanations of this degree of precision and clarity references to atmosphere can amount to little more than the assertion that subjects failed to 'appreciate the structure of the problem' (Hunter 1957b, p. 298) - and even this phrase possesses a clarity which makes it wholly to be preferred to the 'atmosphere' terminology. If this is accepted, then here, as elsewhere, nothing would be lost and something important would be gained if all references to an atmosphere effect were to cease. Perhaps it is significant that Hunter has not subsequently had occasion to use these terms and that quite generally, references to 'atmosphere' in the literature all tend to be of a quasi-historical character referring back to the pioneering studies of Woodworth and Sells where, as I have tried to show, they represent a misunderstanding, at least so far as the 'secondary principles' are concerned.

1.5 The importance of attitudes towards the conclusion of a syllogistic argument. We have already seen, in the monograph by Wilkins, that more errors are made, on the whole, when the truth-value of the conclusion of a syllogism is at variance with the validity of the argument as a whole, than when truth-value and validity are in agreement. Additional support for this contention is to be found in the paper by Janis and Frick (1943) which is noteworthy, in view of some of the papers to be considered later in this section, for the following characteristics: (1) the syllogisms really are syllogisms; they are set out in terms of statements which are easily identifiable as A, E, I or O in the traditional sense; (2) the subjects' attitudes towards the conclusions (in this case simply whether they thought them true or false) were established empirically at the time of the test, and (3) in a way which reduced the risk of these views contaminating, or being contaminated by, the subjects' views about the validity or otherwise of the arguments; (4) there is no reason to doubt the comparability of syllogisms which were supposed to differ only in the truth-value of their conclusions. In these maximally favourable conditions Janis and Frick found that subjects made significantly more errors in judging the validity of valid syllogisms whose conclusions they believed to be false and of invalid syllogisms whose conclusions

they believed to be true than they did in judging syllogisms whose validity agreed with the perceived truth-value of their conclusions. We may with some confidence, therefore, take this to be a well-established fact about syllogistic reasoning.

In the remainder of this section we shall be considering the question whether there is, in addition to this effect, a tendency for errors to be increased by emotional commitments on the subject's part vis-a-vis the content of a syllogistic argument, and in particular by an emotional attitude of acceptance or rejection towards the conclusion. That there is such an effect is something we tend to take for granted; it is also something which is likely to occur in times of national emergency when attempts to develop a reasoned approach to a problem are likely to appear to be defeated by the prevailing atmosphere of emotionality. This is why the two best known papers on the effect of attitudes on the ability to reason syllogistically stem from the period of the Second World War. Both appear to provide evidence for the view that emotionally toned material has a distorting effect on a subject's reasoning processes.

The earlier of the two papers is that by Morgan and Morton (1944). It seems to me to suffer from a number of faults of which the most important are the following. As Henle and Michael (1956) point out, Morgan and Morton failed to establish empirically what their subjects' attitudes towards the conclusions of the various 'emotionally-toned' arguments were. This enabled them, in their interpretations of the shifts which occurred in the conclusions preferred in these arguments as compared with emotionally neutral ones, to 'bend' the evidence to fit their hypothesis, sometimes in a highly unpalatable fashion. A second weakness is that the comparability of the neutral and emotionally-toned arguments is far from being clear in a number of respects. Henle and Michael point out, for example, that the emotionally-toned arguments are very much more prolix than their neutral counterparts. More important, Wilkins's (1928) research ought to have made it clear that a proper comparison between emotionally-toned and emotionally neutral arguments can not be made if the latter are represented, as they are in Morgan and Morton's study, by arguments in 'abstract' form. For Wilkins showed that arguments in this form are generally more difficult than ones couched in 'familiar' terms and that this ^{difference} in difficulty is associated with differences in the kinds of errors made (see p. 7 above).

Most important of all, in my view, so far as the comparability of the two kinds of argument is concerned, the emotionally-toned arguments are presented in terms which encourage a change in the subject's view of the task. Quite generally, these arguments are set out in such a way as to make their relationship to anything which might be called a syllogism highly problematic. The quantity of the premises and conclusions are by no means always clear and in some instances their component statements simply do not state relationships of inclusion between three classes of thing (cp syllogism 7, for example). In these circumstances it is very doubtful if the 'atmosphere' hypotheses of Woodworth and Sells can meaningfully be applied to predict subjects' choices (as Morgan and Morton assume). In any case the authors certainly appear to be mistaken about the 'atmosphere' of the premises of syllogisms 9 and 10 and they introduce a new category of atmosphere to take account of the special characteristics of syllogism 8 which they say 'contains a combination of universal-affirmative and particular-negative atmospheres'.

For these reasons alone one would ^{be} more surprised than otherwise if there were no shifts in the choice of conclusion from abstract syllogism to its emotionally-toned counterpart. There is, however, more to come. Part of the change in the character of the syllogisms from the one half of the test to the other is the inclusion in the premises of the second half of terms such as 'most' and 'usually' which are characteristic of situations in which the subject's task is to assess the probability that the conclusion is true. In addition the conclusions which are supposed to correspond to the I and O propositions of the traditional syllogism are of the form 'X's may be Y's' and 'X's may not be Y's'. This may have at least two different effects on the reasoner: it may reinforce the impression that what he is supposed to do is to assess the likelihood that the conclusion is true, or it may encourage him to take refuge from a more explicit commitment - as Gorden (1953) clearly recognises.¹

For all of these reasons it is impossible to attach any clear significance to the shifts in preference for the various alternative conclusions which Morgan and Morton's research seemed to have uncovered. In particular one cannot accept the authors' contention that these shifts are determined by the subjects' attitudes towards the conclusions of the emotionally-toned arguments.

¹ This paper by Gorden embodies many of the weaknesses of the one at present under discussion (on which it is based). Hardly surprisingly, Gorden finds little evidence of bias attributable to subjects' attitudes.

Lefford (1946) was interested in two things: the extent to which subjects succeeded in distinguishing valid from invalid syllogisms incorporating two different types of material, emotionally significant (E) and emotionally neutral (nE); and the extent to which 'partiality' was shown, i. e., the extent to which subjects judged the conclusions of syllogisms valid to be true and the conclusions of invalid syllogisms to be false. His most important finding must be supposed to be that scores on the half of his test incorporating E materials were considerably lower than on the other, nE half. The usual statistics are not supplied, but graphs show a J-shaped curve in the case of the E syllogisms, with a piling up of scores at the low end of the distribution, and something approaching a normal curve in the case of the nE items. It seems that 74 per cent of subjects scored less than 10 points out of a possible 100 on the E items whereas less than 10 per cent scored as little as this on the nE part of the test. If it were possible to assume that no other difference exists between the items of the two halves, this would represent very strong evidence for the view that a subject's ability to distinguish valid from invalid syllogisms is impaired when he is dealing with syllogisms involving emotionally significant material.

Unfortunately for Lefford's case it is not possible to make this assumption: the two types of item do not differ only in the extent to which they embody emotionally significant material. The fault is not any of the kinds noted in the Morgan and Morton study. Lefford's emotionally neutral items are composed of 'familiar', and not of 'abstract' materials (to use Wilkins's terminology), and, as far as one can judge, they are of the same logical form as their emotionally significant counterparts. The difference lies rather in the proportion of items in the two halves of the test in which the truth-value of the conclusion 'agrees with' the validity of the argument.

As we have seen, Wilkins and Janis and Frick have shown that more errors are made when subjects are asked to judge the validity of arguments whose conclusions are true when the arguments themselves are invalid or whose conclusions are false when the arguments themselves are valid. The assumption is that subjects tend to use the truth-value of the conclusion as a guide to the validity of the argument. Lefford's 'partiality' scores are consistent with this assumption, for he found that 50 per cent of his subjects showed complete partiality, said in every case that the conclusion of an argument was true if and only if

they had already judged the argument to be valid, and that a conclusion was false if and only if they had judged the corresponding argument to be invalid. This was so, clearly to Lefford's surprise, in both halves of the test. If, then, as I have suggested, there are, in the nE half of the test, more valid arguments with true conclusions and invalid arguments with false conclusions than there are in the E half of the test, this by itself would account for the higher scores obtained in the former half.

It is difficult to prove conclusively that this is so. To do so we should really need to know which conclusions the subjects of 1941 believed to be true and which they believed to be false, and Lefford does not include this information in his paper. So far as the nE items are concerned it is possible to be reasonably confident about the views of American undergraduates even of thirty years ago, at least in the majority of cases. The same is not true of the E items, however, for these depend very much on American attitudes to the events and figures of the Second World War. Conscious of the difficulties, however, I have gone over all forty items trying to decide what Lefford's subjects would have been likely to say about the truth or falsity of the conclusions and I believe that this reveals a distinct tendency for truth-value of conclusions to agree with validity of arguments more often in the nE than in the E set of items.

Of the nE items (nos. 1-20 in Appendix D) three (nos. 1, 4, and 6) seem to be of such a nature as to render it nonsensical to ask whether they are true or false. Of the remaining seventeen there are four (nos. 7, 10, 11, and 17) about which I find it very difficult even to guess what the consensus, if any, would have been. Of the remaining thirteen it seems to me virtually certain of twelve and probable of the other one (no. 13) that their truth or falsity would have been judged in the way which makes it agree with the validity of the corresponding arguments. There are, in addition, no items in the nE half of the test which would have been likely to produce errors in subjects influenced in their judgements of validity by their views on the truth-value of conclusions.

In the E half of the test, in contrast, there are six items (nos. 21, 23, 29, 35, 39 and 40) of which it seems to me certain, and five (nos. 24, 25, 31, 32 and 37) of which it seems to me probable that

this effect is operative, and only four (nos. 22, 30, 36 and 38) in which the opposite effect may be supposed, with some degree of probability, to have occurred (the remaining five being conclusions about which I found it impossible to reach a decision). If this is correct, then, as I have already remarked, the relative difficulty of the E items in the test need not be supposed to depend on the emotionally significant material which they incorporate, as Lefford suggests, but simply on the widely differing degrees to which truth and validity, falsity and invalidity, are associated in the two halves of the test.

In reaching this conclusion I was conscious of the possibility of bias on my part and I therefore asked two young American women to go over the conclusions indicating whether they believed that American undergraduates would have regarded them as true or false in 1941. Needless to say, they were not told about the issue involved until after they had completed the task. There proved to be a difference between the two women in the respect that one felt much more confident than the other of her ability to tell how Lefford's subjects would have regarded the conclusions. Considering only those twelve conclusions in the nE half of the test on which the two reached a verdict and agreed on it, ten were thought certainly, and one was thought probably, to be of a truth-value which would have encouraged a correct judgement about the validity of the arguments involved. The remaining conclusion was judged to have a truth-value tending to have the opposite effect.

In the second half of the test their conclusions agree less well with my own - perhaps inevitably in view of the greater difficulties involved, the diffidence of one of my assistants was such that she could not tell how Lefford's subjects would have viewed eight of the twenty E conclusions. Of the remaining twelve only three seemed to both women likely to be viewed in such a way as to encourage the correct view about the validity of the corresponding arguments, whereas three seemed certain and six seemed likely to have the opposite effect. These scores are increased to five, four and eleven respectively if we consider only the views of the more confident of my two assistants.¹ In sum, there is a clear tendency for the nE items to have conclusions which help subjects to reach the right verdict about the validity of the arguments and for the E items to have conclusions which operate in the opposite direction. In these circumstances we must regard Lefford's contention that emotionally significant materials tend to distort a subject's judgement of the

¹ Full information about their judgements is given in Appendix D.

validity of arguments of a syllogistic character as not proven.

A similar sort of weakness appears to be present in the study by Thistlethwaite (1950) which is otherwise notable for its methodological and theoretical sophistication. Thistlethwaite's strategy was similar in some respects to Lefford's. Once again, performance on emotionally neutral arguments - incidentally not of the usual syllogistic variety - was compared with performance on emotionally significant items, the difference between them giving a measure of distortion presumably due to the emotional content of the second set of arguments. An important difference between the studies is that Thistlethwaite worked with two categories of subject, those for whom the emotional arguments really could be assumed to have this character and those for whom this was much less likely to be true.

A further difference between the studies is that the confounding of attitudes towards conclusions (in the sense of beliefs about their truth or falsity) and specifically emotional factors is deliberately built into Thistlethwaite's design: emotional commitments are supposed to be expressed via beliefs about the truth-values of conclusions. Although only two out of the seventy-two arguments used by Thistlethwaite are reproduced in his paper, the whole test, together with introduction, instructions to subjects and key, proved to be available from the Library of Congress Photocopying Service. In his introduction the author explains that the emotionally significant items have been so constructed that their conclusions will appear to be true to ethnocentric individuals in cases where the argument is actually invalid and false where it is valid. Inspection of the anti-Negro items (Appendix D) will serve to confirm that this is indeed the case. Thistlethwaite found that subjects from Southern states where ethnocentrism and, in particular, anti-Negro sentiments, are apparently prevalent showed much greater distortion than subjects from Northern states where this is not the case. The difficulty is, as I see it, to separate the established effects of beliefs about the truth-value of conclusions from the effects, on one's ability to tell a valid argument from an invalid ^{specifically} one, of emotional commitments.

We come finally, and rather briefly, to a paper in which a claim is made to show that, at least in one particular case, emotional commitment did not interfere to any perceptible extent with the ability of subjects to distinguish valid from invalid arguments. This is the paper by Henle and

Michael (1956) already referred to in this section. Once again the evidence is equivocal. Two groups differing in their attitudes towards Russia showed no significant differences in their views as to which conclusions could be inferred from premises 'concerned with communism, Russia or related matters'. On the other hand, as Henle and Michael admit, the differences between the two groups in attitude were relatively small: too few subjects declared themselves to be pro-Russian for statistical treatment of the results to be possible, so that the comparison was between subjects who declared themselves to be anti-Russian and subjects who professed to have no strong feelings, one way or the other, about Russia. Henle and Michael add one other point from their results which seems to them to be inconsistent with the view that arguments with emotionally significant contents are harder to judge correctly than arguments without such contents. This is that subjects scored higher on the Russian syllogisms than on the 'abstract' syllogisms which were supposed to act as a point of comparison. They comment: 'We do not have strong attitudes to X's and Y's and Z's.....' The point appears to me to be a debatable one, but even if it were allowed to be true, it would not serve to support the conclusion which Henle and Michael wish to draw in the absence of any proof that emotional significance is the only relevant respect in which the two sets of syllogisms differ.

In this section we have reviewed evidence which seems to lend unequivocal support to the view that attitudes towards the conclusion of an argument play an important part in determining the success with which a person evaluates its validity. However, this is true only if by 'attitudes' we mean to refer to the subject's views about the truth or falsity of the conclusion. None of the attempts to establish a separate role for emotional factors which we have reviewed, can be regarded as successful, although the objections to the study by Thistlethwaite must be regarded as tentative in character.

If the conclusions of Section B of the Valentine test were such as to be clearly true or clearly false, it would be necessary to consider the possibility that such individual differences as we have found between subjects of equally high ability and attainment are due to a greater proneness, on the part of the poor reasoners, to be affected by this factor, and we should then, as a first step, have had to establish of how many arguments in this test it was true that the validity of the argument 'disagreed' with the truth-value of the conclusion. It is a point which

one might hope to establish in a relatively straightforward empirical fashion. In actual fact an examination of the conclusions of the arguments in the two forms of Section B of the Valentine test¹ makes it clear that, at least as they stand, many of these are not such that one can very sensibly ask whether they are true or false. Generally speaking, these are the items in which the conclusion refers to some particular thing or class of things with which the subject could not be expected to be familiar - if only because they are imaginary (the people of Tutland or Abasiland in items 15A and B) or unspecified (the prisoners of W9) or indefinite (the John Smith of V5). In one or two cases it seemed possible to translate the conclusion, without drastically altering the point at issue, into a form in which it did make sense to ask whether it was true or false. Thus the conclusion, 'This (seditious) pamphlet must be suppressed' was presented in the form, 'Seditious pamphlets must be suppressed' - though even here it might be suggested that the original conclusion commits one to no more than the view that some seditious pamphlets should be treated in this way.

Putting these considerations into practice, a sheet was prepared on which thirteen of the original twenty-four conclusions were set out together with the following instructions:

The following are statements appearing as conclusions of arguments in Valentine's Reasoning Tests. Because an important factor in determining whether a person reasons correctly or not appears to be the truth or falsity of the conclusions of the arguments he is considering, it is very important to know whether the following statements appear to most students (for whom the Valentine test is intended) to be true or not.

Look at each statement carefully and try to decide whether it is true or false. In some cases you may not be able to make up your mind one way or the other. In such cases just write a question-mark in the left-hand margin. If you think the statement is definitely true or definitely false, write T or F respectively. If you think it is probably true or probably false, write T? or F? respectively.

Sixty members of a first year psychology class subsequent to the one involved in the testing sessions referred to in Appendix A of this thesis were asked to complete the task described in the above instructions (which were also read out to them). The frequency with which each of the thirteen conclusions was ascribed to the five categories of truth and

¹ See Appendix B where the two forms, V and W, are reproduced.

falsity is shown in Table 1.2 below.

TABLE 1.2

FREQUENCY WITH WHICH CONCLUSIONS OF ARGUMENTS IN SECTION B OF THE VALENTINE TEST WERE SAID TO BE TRUE OR FALSE (N = 60)

Item No.	T	T?	?	F?	F	Conclusion predominantly:	Argument:
V6	22	18	1	3	16	T(?)	Valid
V7	--	1	10	11	38	F	Invalid
V8	11	19	1	6	23	?	Invalid
V10	4	8	4	9	35	F	Invalid
V12	5	2	1	3	49	F	Valid
V13	4	2	2	5	47	F	Valid
V14	10	10	2	10	28	F(?)	Invalid
W6	1	1	--	7	51	F	Invalid
W7	--	--	--	3	57	F	Valid
W10	4	12	1	6	37	F(?)	Invalid
W12	10	11	--	8	31	F(?)	Valid
W13	5	13	--	7	35	F	Invalid
W14	2	1	5	7	45	F	Valid

To the right of the table I have tried to indicate whether the conclusions were on the whole deemed by these sixty subjects to be true or false. It will be seen that all but two were regarded, with varying degrees of unanimity, to be false. More important, a comparison of the truth-values of the conclusions, as thus determined, with the validity of the arguments suggests that there are five items in which a subject who is influenced in his judgment of the latter by his view about the former may be more likely to make a mistake than a subject not influenced in this way. These are items V12, V13, W7, W12 and W14. In fact, as will be seen in the next chapter, there were large differences in the extent to which 'good reasoners' and 'poor reasoners' responded correctly to some of these items, and it may be held likely, therefore, that the differences in reasoning ability in which I am interested are to be explained at least partly in terms of the operation of this factor. How large a part of the explanation is to be found here it is, of course, difficult to say with any certainty. If it is correct to assume, however, that a subject's views about the truth-value of the conclusions of the arguments

is likely to militate against success in the logical reasoning task only in five out of the twenty-four items of the two forms, it is very likely, at least, that other factors are operative.

1.6 Some more general considerations To establish that an argument is likely to be regarded as valid if its conclusion is true and invalid if its conclusion is false is not, of course, to explain how this happens. Perhaps the most likely explanation (Richter, 1957) is that the subject confuses the task of saying whether an argument is valid with that of saying whether the conclusion is true, fails to notice that the latter is not identical with the former or, more likely, fails to realise that the truth of the conclusion does not entail the validity of the argument or, except in the special case where the premises are true, the falsity of the conclusion the invalidity of the argument. Henle and Michael (1956) and Henle (1962) offer evidence in support of the view that, in at least some cases, subjects 'do not accept the logical task', do not, that is, maintain a clear distinction between questions about the validity of arguments and questions about the truth-value of conclusions.

In the first experiment described in the 1956 paper Henle and Michael replicated, more or less, the Morgan and Morton study described in the previous section. The outcome was very much in line with that reported by Morgan and Morton. In a second experiment an attempt was made to retain the 'emotionally significant' character of the non-symbolic syllogisms while simplifying them and making them less cumbersome - by using statement forms more like the orthodox A, E, I and O ones and by using terms which consisted mostly of one or two words instead of the long phrases of Morgan and Morton. The result, with a different group of subjects, was an increase in the percentage of correct responses from 37 (using the Morgan and Morton material in the first experiment) to 56 (using the simplified, but still 'emotionally significant' arguments) and rather surprisingly, though not discussed by the authors, an increase from 26 to 56 percent in the symbolic syllogisms.

The main purpose of this second experiment was, as we have seen in the previous section, to throw doubt on the contention that emotional subject-matter presents special difficulties in a logical reasoning task. At the same time Henle and Michael represent it as first step in the 'progressive clarification' of the logical task. In view of this the following comments would appear to be in order. In the first place, the fact

that there was, as we have just noted, a large improvement in performance on the syllogisms in symbolic form, from the first to the second experiment, suggests that other factors were operative in addition to any 'clarification of the task' achieved by the simplification of the contents and the regularisation of the form of the 'emotionally significant' syllogisms. In particular, in view of Wilkins's contention that it is easier to recognise valid syllogisms as such rather than invalid ones, it is important to note that 12 of Henle and Michael's 20 syllogisms were valid whereas those used by Morgan and Morton were predominantly invalid. Secondly, though the authors say that instructions in the experiment were minimal, some clarification of the task probably occurred as a result of the fact that the task was more readily identified as one calling for deductive reasoning: the form of Morgan and Morton's arguments, as we saw in the previous section, encouraged the belief that the task was one of assessing probabilities. Thirdly, the changes in the content of the 'emotionally-toned' syllogisms may have made ^{their} subjects' task easier but cannot reasonably be said to have done so by 'clarifying the task': it is much more plausible to represent the difference between the syllogisms of the first and second experiments as being of the same general kind as the difference between the Wilkins syllogisms couched in 'familiar' terms and those couched in terms of 'tiksatopses' and other linguistically unassimilable terms.

In their third experiment Henle and Michael tried to achieve a further 'clarification of the task'. 'This experiment employed rather full oral instructions on how to solve syllogisms. Subjects were shown how to use diagrams in solving syllogisms, and several examples were worked through. When the experimenter had finished her exposition, subjects were allowed to ask questions; and not until the task was entirely clear did the experiment proceed.' (Henle and Michael, p. 125) The result, with a third, and much smaller, group of subjects, was that the proportion of correct solutions went up from 56 per cent in Experiment II to 82 per cent in Experiment III on syllogisms incorporating emotionally significant material and from 49 to 83 per cent on syllogisms in symbolic form.

Unfortunately, it is still not clear how far the procedure described can really be said to have 'clarified the task' and to have done no more than that, or how it relates to the view that subjects fail to understand the task in the sense that they 'do not distinguish between

a conclusion that is logically valid and one that is factually correct or one with which the subject agrees' (Henle, 1962, p. 370). Henle and Michael's 'clarification of the task' as described above seems to have included the demonstration of techniques for establishing the validity or invalidity of an argument and also some actual or vicarious practice in using these, and perhaps some other, less formal, techniques. Of course subjects may have acquired an understanding of the task as a more or less accidental by-product of the demonstration and practice, but the understanding of the task is separate from, and strictly speaking, presupposed by an ability to use the techniques referred to. Euler's circles and Venn's diagrams were originally introduced to make it easier to complete a task whose character was assumed already to be clear, and in so far as it was the use of such devices which accounted for the improved performance in Henle and Michael's third experiment it is not obvious that this represents a greater understanding of the task itself.¹

Henle's later paper (1962) is supposed to show that the gap between the thought processes required by logic and those which normally occur - the way we ought to think and the way we actually do think - is less than research of the kind reviewed in the previous section might lead one to suppose. Errors in syllogistic reasoning tasks, Henle suggests, are due less to illogical reasoning strictly so-called than to factors of which 'failure to accept the logical task' is one. The others she mentions are the re-statement of a premise or conclusion so that the intended meaning is changed, the omission of a premise, and the slipping in of additional premises.

No quantitative results are offered for 'as many authors have shown, the incidence of error in deductive reasoning depends on the form of the syllogism and its contents as well as on instructions to subjects. Quantitative results would have relevance to the particular conditions studied here, whereas an inquiry into the nature of the errors obtained might be of more general interest' (ibid., p. 370). In fact the arguments used are so different in form from the recognisably syllogistic character of the arguments used in the study described in the 1956 paper

¹ There must be a serious doubt, also, as to the comparability of the samples used in the second and third experiments. In the former the subjects were two classes from New School (66 in all) plus one class from Hunter College (34 subjects); in the latter there were only 15 subjects, all of them graduates enrolled in a class on public speaking.

that it is difficult to be sure whether Henle thinks - or is right to think - that the errors obtained with the latter type of material are also likely to occur in the other, more formal, kind.

An example will make the difference clear. 'Syllogism 6' reads as follows:

A group of women were discussing their household problems. Mrs Shivers broke the ice by saying: 'I'm so glad we're talking about these problems. It's so important to talk about things that are in our minds. We spend so much of our time in the kitchen that of course household problems are in our minds. So it is important to talk about them.' (Does it follow that it is important to talk about them? Give your reasoning.)

The trouble with such an informal presentation of the task is that it positively invites subjects to consider the broader question whether the conclusion is established by the argument rather than the question whether the conclusion must be true if the premises are true. The former question is the one we may have to decide in everyday situations, when someone is trying to convince us of the truth of some statement, but it is not the question posed by normal deductive reasoning tasks of the kind discussed in this chapter. The point is an important one because a conclusion can be said to be established by an argument only if the argument is valid and the premises are true. Hence one can normally refuse to accept the conclusion of an argument either on the grounds that the argument is invalid or on the grounds that one or more of the premises are false. It is hardly surprising, therefore, that some of the examples Henle offers of responses which show the subject 'failing to accept the logical task' are cases where the subject questions the truth of a premise. Examples which do not fall into this category seem to consist of (a) an acceptance of the argument and (b) the suggestion of additional reasons for supposing the conclusion to be true.

In most deductive reasoning tasks, as I have just been implying, the subject is clearly required to consider the validity of the argument and simply to assume the truth of the premises. In the Valentine test the instructions of Section B include the following points:

You must assume first that the given premises (i.e. the statements underlined) are true. The problem is, in each case this: granted that these statements are true, is the other statement (the conclusion of the argument) necessarily true?

REMEMBER THAT YOU MUST ASSUME THAT THE UNDERLINED STATEMENTS ARE TRUE

The underlining and the capitals are, of course, in the original. It is possible, obviously, that despite the heavy emphasis thereby laid on the relevant distinction between the questions of validity and proof, subjects proceed to ignore it or, more likely, eventually to lose sight of it. Some evidence relating to the point will be presented in the next chapter. In the meantime it is clear that the examples presented by Henle can scarcely be said to establish the likelihood of such an eventuality since, in strong contrast to the Valentine, her test seems to have gone to no trouble to make the relevant distinction clear. In general, as she herself admits, in the quotation from page 370 and elsewhere, the kinds of mistake made by her subjects are to be ascribed, to an unknown extent, to special characteristics of the 'syllogisms' they were asked to judge. Thus, 'the informal manner in which the premises were set out' made it possible for subjects to omit premises, consciously or otherwise. Equally, it is sometimes difficult to be sure how to quantify the premises of the sample syllogisms she presents and it would therefore be relatively easy for subjects to 'restate' them without their being aware that they were doing so. In general, then, though mistakes of the kinds Henle describes do occur in circumstances which favour them, it is far from certain that they will occur with any considerable frequency in the rather formal kind of deductive reasoning task with which this chapter, and indeed this thesis, is concerned.

1.7 Summary of the main points This chapter has attempted an evaluation of the literature relevant to the problem of this thesis which may appear to the reader to be too long, and critical in the strong sense of finding much to complain about. The only defence that can be offered is that the positions allegedly established in the papers reviewed are mostly so well entrenched in psychology that nothing less than a major effort seemed adequate to expose the inadequacy of the methodological and/or theoretical basis on which they rest.

Contributions which largely escaped criticism were the monograph by Wilkins (1928) and the short paper by Janis and Frick (1943), both of which seem to me to have established that it is more difficult to judge the validity of an argument if the truth-value of the conclusion 'disagrees with' - or is believed by the subject to disagree with - the validity of the argument. The probable effect of this factor on performance on the Valentine test, Section B, was considered at the end of section 1.5 and was thought to be real but minor. Wilkins's monograph was also welcomed

for its evidence that some undergraduates of above average ability do have serious difficulties with deductive reasoning tasks and that subjects frequently illicitly 'convert' an A proposition, and for its points about the relative difficulty and discriminating power of different formal fallacies, reference once again being made, in the light of these, to the arguments of the Valentine test.

The discussion of the 'atmosphere effect' (Woodworth and Sells, 1935, Sells, 1936) attempted to demolish the, by now traditional, assumption that this effect accounts for errors in syllogistic reasoning in terms of a factor which is at least non-logical and perhaps actually contra-logical. It was argued that the effect in question is of small importance as an explanatory device where it seems best to apply - in accounting for the tendency of subjects to convert A, E, I and O propositions. In its later role of explaining why subjects tend to prefer one invalid conclusion of a syllogism to other invalid conclusions it was criticised, not for failing to do this on the basis of its 'secondary hypotheses' - on the contrary, and despite Chapman and Chapman (1959), it seems to be rather successful in this respect - but for its misidentification of the factors probably involved. In particular, it was shown that only an arbitrary criterion of the 'similarity' of premises to conclusion could make this seem the operative factor and that the important considerations are probably logical ones: subjects susceptible to the atmosphere effect (i.e., who show the preferences, as between invalid conclusions, predicted by it) are being guided in their choice of conclusion by some, but not all, of the 'rules of the syllogism'. (Sells's own 'paradoxical' finding that 'bright' subjects are more susceptible to the effect than 'dull' ones lends credence to this interpretation.) More generally, it was suggested that no good grounds have been presented for the retention of references to 'atmosphere' in attempts to account for failures in a problem-solving task.

I tried to show that there is no good evidence to support the widely held view that the subject's attitude towards the conclusion of a syllogism - other than his belief that it is true or that it is false - is partly responsible for the adequacy or otherwise with which he judges its validity. All attempts to establish that a subject's judgement is distorted by his emotional reactions to the subject-matter of the arguments he is called upon to evaluate were shown to have failed to control for some other variable, notably the 'agreement' or 'disagreement' of the validity of the argument with the truth-value of its conclusion. On the

other hand, the attempt by Henle and Michael (1956) to establish that attitudes towards the conclusion have no effect on a subject's ability to distinguish valid from invalid arguments is also methodologically weak.

Other possible sources of error in syllogistic reasoning tasks were finally considered, including what is probably the most plausible, a failure to understand, or to adhere to, the requirements of the task. Here, too, serious doubts were expressed about methodology or about the relevance of the conditions described in papers purporting to establish the operation of such factors to those obtaining in the criterion Valentine reasoning test.

CHAPTER TWO

SELECTION OF SUBJECTS, PRELIMINARY EVIDENCE, AND THE 'FIVE TYPES OF STATEMENT' TASK, PART ONE: DESCRIPTION OF THE MATERIALS AND PROCEDURE

Summary The procedure used to establish two groups of undergraduates of comparable academic ability and attainment but of widely different deductive reasoning ability is described. Evidence about differences between members of the two groups which might bear on the difference in deductive reasoning ability was drawn from patterns of error in the criterion test and from answers to a questionnaire and is reviewed in this chapter as a preliminary to the more extended, experimental study of differences between the groups which is the subject of the remainder of this thesis. Finally, the details of the materials and procedure used in the first experimental task, the 'Five Types of Statement' task, are set out.

2.1 Selection of subjects I have referred on several occasions in the previous chapter to Section B of Valentine's Reasoning Tests as the criterion by which I proposed to judge the syllogistic reasoning ability of my subjects. The nature of this Section and the fact that there were two forms of it at my disposal (one of them devised by myself¹) will have become apparent. I have also remarked that it comes closer than any other published test to the kind of task used in the Logic Department of Glasgow University for the teaching and assessment of the ability to distinguish valid from invalid arguments.² In November 1966 the entire membership of the Ordinary Psychology class at Glasgow sat both forms of the test on succeeding Fridays and Terman's concept Mastery Test on the intervening Tuesday. The results of this testing programme not only made it possible to carry out the technical evaluation of the published form of the Valentine test presented in Appendix A but also, along with information about performances on the Higher Grade of the Scottish Certificate of Education, furnished me with data which could be used to select subjects for the experimental study which is the concern of the remainder of this thesis.

To this end a composite academic ability and attainment ('AAA')

1 See Appendix A, Sections A.2 and A.3, and Appendix B, where both forms of the test are reproduced.

2 I have also mentioned various weaknesses of this test which became apparent only after its use for the present purpose. These are described in Appendix A, Sections A.4 - A.6, it being suggested, at the end of that Appendix, that these weaknesses do not seriously affect their usefulness as a selection instrument for present purposes.

score was calculated for each subject,¹ based on his scores in Section A of the Valentine test, in the Concept Mastery Test (adjusted to allow for a pronounced inter-faculty difference in favour of Arts students²) and on the Scottish Certificate of Education, Higher Grade³ - 'scores' in the last case being found by awarding 3 points for an A pass, 2 for a B and 1 for a C, a procedure also adopted by Nisbet and Napier (1970). The resulting distributions are presented in Appendix E. Scores from these three sources were weighted, respectively, in the proportions 1 : 3 : 4, the heavy weighting on the S.C.E. component being intended to reflect the amount of information about the subjects' educational attainments, as opposed to ability, represented by these scores. Finally, for convenience of comparison, the AAA scores and the scores on the two forms of the Valentine test, Section B, were transformed into T-scale equivalents with, of course, a mean of 50 and a standard deviation of 10.⁴

Using these T-scale scores on the two measures I constituted my two experimental groups as follows. Members of the 'poor reasoners' (PR) group had to be at least average on the AAA measure but below average on the Valentine, Section B: in fact I set an arbitrary minimum difference between the two scores of 10 T-scale points, equivalent, of course, to one S.D. on either measure. For each student who satisfied these criteria I tried to find a person of the same faculty and sex with the same AAA score but with a Valentine score within a point or two of his AAA score, in other words, of the same, at least average, academic ability but without the relative weakness in deductive reasoning ability. The outcome was, of course, that the difference between the Valentine, Section B scores of any pair of subjects was in no case less than 10 T-scale points. I regarded identity of sex and faculty of lesser importance so long as the difference between members of a pair was confined to only one of these factors.

1 Of the 320 members of the class 286 had scores on all three measures, the main source of loss being students whose entrance qualifications were in G.C.E. or were otherwise difficult to compare with S.C.E. Higher passes.

2 The adjustment took the form of finding standard scores for students from the two faculties based, of course, on the relevant means and S.D.s.

3 Counting passes gained in the fifth year at school only: to have included sixth year passes as well would have been to introduce an imponderable, for although every Scottish school child who proceeds to University has a fifth year at Secondary School, some do not stay for a sixth year and, of those who do, some try to improve their grades on subjects already passed, others try to pass on a subject for the first time and, of course, others regard their sixth year as an opportunity for engaging in activities of an altogether different kind.

4 See Guilford (1965) p. 518ff.

Altogether there were 26 pairs of students meeting these criteria. Of these, however, only 22 pairs actually completed the tasks included in the first stage of my research. Means and standard deviations (in T-scale units) for the two groups, poor reasoners (PR) and good reasoners (GR), are given in Table 2.1, individual scores being relegated to Appendix E. The mean and standard deviation for the original pool of 286 were, respectively, 50 and 10 for both AAA and V.R.T. Section B.

TABLE 2.1

MEANS AND STANDARD DEVIATIONS OF AAA AND V.R.T. SECTION B SCORES FOR THE TWO EXPERIMENTAL GROUPS: POOR REASONERS (PR) AND GOOD REASONERS (GR)

		<u>A A A</u>		<u>V.R.T. SECT. B</u>	
	N	Mean	S.D.	Mean	S.D.
GR group	22	59.36	5.17	60.00	4.50
PR group	22	60.18	5.60	42.82	4.79

As we shall see, the overwhelming trend of the experimental data is to confirm the belief that the large difference in mean V.R.T. Section B scores between the two groups reflects a real difference in their ability to cope with tasks calling for deductive reasoning ability or its components; in none of the experimental measures did the PR group as a whole excel over the GR group - though the differences between the groups reached statistical significance on only a few occasions.

So far as the adequacy of the AAA score as a means of securing equivalence of the two groups in academic ability and attainment is concerned, the only subsequent check on this was their scores in the two class exams in psychology, the inadequacy of this as anything more than a rough guide being fully acknowledged. Table 2.2 presents data for both groups and for the Ordinary Psychology Class as a whole. It shows that the PR group was, if anything, superior to the GR group on both examinations. On the other hand, the mean score of the GR group is not only lower than that of the PR group but obviously not significantly higher than that of the class as a whole - so that this evidence, for what it is worth, provides only partial support for the contention that the two groups are of equal and above average academic ability and attainment.

TABLE 2.2

MEANS AND STANDARD DEVIATIONS OF THE PR AND GR GROUPS AND OF THE WHOLE
ORDINARY PSYCHOLOGY CLASS IN TWO CLASS EXAMINATIONS

	N	1st Class Exam		2nd Class Exam	
		Mean	S.D.	Mean	S.D.
PR group	21*	60.95	11.57	55.29	8.27
GR group	21*	54.33	9.15	51.62	9.58
Whole class	311	55.97	10.75	49.65	10.60

* One member of each group failed to sit the exams.

2.2 Some preliminary evidence Having established my two experimental groups it seemed appropriate to begin by reviewing the evidence already at my disposal about the differences between them. In particular it seemed appropriate and worthwhile to study their answers to the questionnaires which all members of the class had completed at the end of the testing programme and to look for differences in the patterns of errors of the two groups in the criterion test of deductive reasoning ability.

The questionnaires offered evidence about a number of things of potential relevance to the point at issue - not only about the comparability of the two groups on age, level of academic aspiration and logical training and interest, but also about their reactions to the Valentine test and their views about their competence in the kind of task under consideration. Some of this information might suggest additional points of departure in the search for factors underlying individual differences in this ability, others could only provide interesting information, for example, about the extent to which the PR group's relative inferiority in this area made itself felt. The relevant questionnaire items and the distributions of answers for the two groups are given in detail in Appendix E.

For most of the items the distributions are obviously not significantly different for the two groups. This is true of sex, age, faculty (Arts or Science), year of study, and whether or not aspiring to an Honours degree. It is also true of answers to questions about attendance at classes on logic, on logic books read, and on the use of logical aids in the completion of Section B of Valentine's test. Roughly the same proportion of each group said the task in Section B was clear or very clear,

and that they had had to resort to guessing or to looking at the reasons offered before deciding whether the argument was valid. Just about as many from each group found the Concept Mastery Test more enjoyable than the Valentine, found the second form of the Valentine they did easier - and for the same natural reason, viz., practice. Similarly, 9 subjects in each group mentioned a factor which might have accounted for their doing less well on a test (not always specified) than they might otherwise have done.

Noticeable but not significant differences in trend were apparent in answers to the questions whether the Valentine test seemed a good test of all-round ability (the good reasoners, surprisingly, being less well disposed towards it), whether the subject thought he was good at spotting weaknesses in other people's arguments, and whether the mere existence of a time-limit prevented him from concentrating on the task in hand. The one significant difference ($\chi^2 = 9.715$, $p < 0.005$) was on the question whether the subject thought he could have done better on the Valentine test if he had had more time. As one might have expected, the poor reasoners were significantly more liable to answer this question in the affirmative, estimates of the amount of extra time required being mostly of the order of 15 minutes (as with the GR subjects who said they could have done better with more time) though three PR subjects thought they could have used 30 minutes more, and one subject said he would have needed another hour to do himself justice. Of the 17 PR subjects who said they needed more time 3 said they would have spent it on Section A, 5 on Section B and 9 on both. The corresponding figures for the GR group are 2, 3 and 1.

Extreme slowness was a feature of some of the PR subjects in my previous, pilot study. Its significance is a little difficult to interpret if we do not simply identify slowness with incomprehension - as, of course, one might in this kind of task. Latency of response has commonly been taken as a measure of difficulty - as it was, indeed, in one of the tasks included in the first stage of the experimental work described in this thesis. In the previous research referred to, on the other hand, PR subjects were able to achieve very high scores on the Advanced Matrices Test (Raven, 1965) provided they were given an unlimited amount of time, spread over two sessions (Wallace, 1965). This is hardly consistent with the 'incomprehension' hypothesis except,

perhaps, stated in terms of speed of comprehension.

With respect to the other source of information about my subjects already at my disposal, namely, their patterns of error in the two forms (V and W) of Section B of the Valentine test, it seemed at least possible that an inspection of the items in which the largest discrepancies occurred might suggest where the sources of the differences between the groups lay or tentatively support the possibilities suggested by the literature. To this end I tabulated them as in Table 2.3 below, in which a distinction is made between success on both tasks (identifying an invalid argument as invalid and also a statement of the flaw in such an argument) and success on only the first. Where an argument is valid, there is only the first.

TABLE 2.3

NUMBER OF SUCCESSES ACHIEVED BY PR AND GR GROUPS IN VALENTINE B ITEMS

Items:	V5*		V6		V7		V8		V9		V10		V11		V12	
Level of Success	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR
Both tasks	18	22	--	--	11	17	13	18	--	--	16	20	13	20	--	--
First task only	0	0	8	19	8	3	4	4	16	20	2	1	3	1	12	19
Neither	4	0	14	3	3	2	5	0	6	2	4	1	6	1	10	3

V13		V14		V15A		V15B		W5		W6		W7		W8		W9		W10		W11	
PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR	PR	GR
--	--	5	9	4	8	2	10	21	22	7	14	--	--	15	20	--	--	13	22	5	15
17	21	5	1	3	4	15	8	0	0	12	7	9	14	4	2	20	22	6	0	9	4
5	1	12	12	15	10	5	4	1	0	3	1	13	8	3	0	2	0	3	0	8	3

W12		W13		W14		W15A		W15B	
PR	GR	PR	GR	PR	GR	PR	GR	PR	GR
--	--	4	12	--	--	6	12	5	17
14	20	3	2	11	20	1	3	8	3
8	2	15	8	11	2	15	7	9	2

* I.e., Form V, item 5.

I began by testing the distribution of scores for the two groups for each item to see which differences, if any, were significant. In fact, only four were: V6 ($\chi^2 = 9.586$, with Yates correction, $p < .01$),

V12 ($\chi^2 = 3.931$, Yates correction, $p < .05$), W14 ($\chi^2 = 6.972$, Yates, correction, $p < .01$), and W15B ($\chi^2 = 13.273$, $p < .01$). Undoubtedly the most interesting aspects of this outcome are (1) that all the arguments involved were valid and (2) that the second and third are alternative forms of a single argument. The fact that W15B is a member of this group may reasonably be taken to reflect or confirm the PR group's need for greater time, for, of course, this is the last item in form W. Responses to 15A and 15B are also difficult to interpret because of a change in the nature of the task which occurs in these items. (On this point see Appendix A, Sect. A.4.)

Items V12 and W 4 are of the form: "No X's are Y's. Only X's are Z's. Therefore no Z's are Y's." Corresponding reasons for supposing the two arguments to be invalid were selected by a majority of PR subjects (the same reason favoured by members of the Ordinary Psychology class as a whole), viz., (i) in V12 and (iii) in W14 (the numbers choosing the four reasons being, in order from (i) to (iv), 6, 2, 0 and 1 (with one blank) in V12 and 0, 2, 6 and 2 (with one blank) in W14). The reason in question offers a counterargument purporting to show that some Z's are Y's. A variety of comments and interpretations are possible. Henle (1962) would no doubt say that subjects who choose this reason 'fail to accept the logical task', do not, that is, restrict themselves to the question whether the conclusion must be true if the premises are true. More positively, one might suggest that these subjects confuse this 'logical' task with the task of saying whether the conclusion is established or proved. The conclusions of both arguments are in fact believed by a majority of students (see p. 33 above) to be false and one can imagine a subject who makes the response in question arguing that there must be something wrong with the premises since they appear to prove something which is manifestly not true. In doing so, of course, such subjects would indeed be failing to accept, or to adhere to, the terms of the task as set out in the instructions, and inasmuch as this happens more often with PR subjects than with their GR counterparts, it does perhaps lend some support to the view that the difference between these two groups is at least partly due to a tendency on the part of the former to lose sight of the 'logical task'.

In some ways the situation with regard to V6 is easier to interpret. This item is of the form: "Everything is either an X or else aY. No Y's are Z's. Therefore only X's are Z's." Considering the four

reasons offered, the numbers of PR subjects choosing each were, again in order from (i) to (iv), 1, 1, 1, and 8 (with three blanks), the corresponding figures for the original group being 17, 14, 9, and 46. Reason (iv) purports to detect an inconsistency in the argument, contending that some X's are Y's (as, indeed, the first premise allows) and, therefore, according to the second premise, not Z's. What gives the selection of this reason its special interest is the fact that this is a competent objection if, and only if, the conclusion is taken to mean or imply that all X's are Z's, a misunderstanding of the 'Only X's are Y's' form of statement which is the mirror-image¹ of the confusion about the meaning of the universal affirmative the importance of which we saw in the last chapter Wilkins had stressed. Finally, it ought perhaps to be mentioned that the corresponding reason is favoured by a majority of the PR subjects who get the corresponding W item 7 wrong. In this case, however, the difference between the two groups fails to reach significance on the χ^2 test, largely because so many of the GR group took the same view of the argument - six of the eight who thought the argument invalid selected the same reason.

It is, of course, possible to cast one's net rather wider in the search for clues as to the sources of the differences between the two groups - for example, by ranking the 24 items in terms of the size of the discrepancy between the success rates of the two groups on each and then scrutinising those items in which the discrepancies are greatest. If, for example, one adds the number of successes in any item on the first task only (determining the argument's validity) to twice the number of successes on both tasks (in items where there are two tasks) and then calculates the ratio of PR to GR 'scores' obtained in this way, the result is as set out in Table 2.4 below.

TABLE 2.4

RATIO OF PR TO GR SCORES ON THE TWENTY-FOUR ITEMS OF VALENTINE'S TEST,
SECTION B, FORMS V AND W

Item	V6	W13	W15A	W15B	V15A	W14	W11	V12	W7	V15B	W12	V11
PR/GR	.42	.42	.48	.49	.55	.55	.56	.63	.64	.68	.70	.71
	W10	W6	V8	V14	V9	V13	W8	V7	V5	V10	W9	W5
	.73	.74	.75	.79	.80	.81	.81	.81	.82	.83	.91	.95

1 Almost literally since in logic 'Only S is P' is taken to mean 'All P is S'.

The difficulty, of course, is to decide which of these differences are large enough for it to be worth while examining them for clues. There is, to my knowledge, no non-arbitrary way of doing this. It is, however, noteworthy that the first ten include the four items whose distributions proved to be significantly different for the two groups on the χ^2 test, together with their alternative forms (W7 and V15B). Also included are the pair V15A and W15A and two 'odd' items, W11 and W13. The presence of the 15A's as well as the 15B's must be taken to confirm the relative slowness of the PR group - 4 of them, as compared with 0 of the GR group made no response at all to at least one of these items, and 7, as compared with 2, failed to mark one of the reasons. The special character of these four items must also, of course, be borne in mind.

As to W11 and W13, the fallacy in the former of these is that of supposing the contradictory of a universal affirmative to be the corresponding universal negative. 14 PR subjects recognised that the argument was fallacious; of these, however, only 5 selected the correct reason, 8 of the remaining 9 selecting a reason which casts doubt on one of the premises. Inspection of the 'correct reason' shows it to be rather unsatisfactory in the sense that it fails to make the exact weakness in the argument clear. With this possible complication in mind, however, responses of the PR subjects to W11 may be construed as additional evidence that these subjects are more prone than their GR counterparts to lose hold of 'the logical task' and to substitute for it the question whether the conclusion is proved by the premises - for in order for that to be achieved the premises have to be true as well as the argument valid. In W13 the argument would appear to be valid to anyone who supposed (as in V6 and W7) that 'only X's are Y's' means or implies that all X's are Y's. (The argument is of the form: "Only X's are Y's. (All) Z's are X's. Therefore (all) Z's are Y's." This is a syllogism in Barbara if the first premise is read as 'all X's are Y's'.) Only 7 of the PR group realised that the argument was fallacious and of these only 4 selected the reason which clearly states the mistaken reading referred to above. In contrast 12 of the 14 GR subjects who said the argument was invalid selected this reason. This appears to me to be additional evidence for the view that PR subjects have a poorer grasp of the meaning of statements of the form 'Only X's are Y's' than their GR counterparts and, in particular, are more likely to assume that it implies that all X's are Y's.

Taking all the evidence at my disposal - that is to say, from

previous work on syllogistic reasoning, as well as the data presented in this chapter - it seemed possible that PR subjects differed from their GR counterparts in any or all of the following respects:

- (1) they have a less secure grasp of the meaning of statements of certain logically important types, notably statements of the forms 'All S is P' and 'Only S is P', both of which play an important role in the arguments of Section B of the Valentine Reasoning Tests;
- (2) they have a less secure grasp of the requirements of what Henle calls 'the logical task';
- (3) they are (perhaps for that reason) more likely to be influenced in their judgement of the validity of the arguments by their beliefs about the truth or falsity of the conclusions;
- (4) perhaps because of the uncertainties implied in all of the above they appear to need more time to complete a logical reasoning task than their GR counterparts.

The remainder of this thesis will describe some experiments intended to throw further light on the first of these possibilities.

2.3 The 'five types of statement task': rationale and procedure

Having raised the question whether there may be some difference between PR and GR subjects in the completeness of their understanding of certain types of statement, one is immediately faced with the problem of how to determine, with a sufficient degree of sensitivity, whether a person does fully understand a particular type of statement. Philosophers as well as psychologists have spent a considerable amount of time debating this question¹ or, more usually, at least as far as philosophers are concerned, the intimately related question of the criteria we should employ in deciding exactly what a statement of a certain type means. Out of all this discussion one thing at least seems to emerge - namely that a person who understands a statement can draw certain inferences from it if it is true - or, to reformulate this in the preferred terms of statement-meaning, a statement's meaning can be explained - at least partly - in terms of the other statements which it implies. In the case of statements whose distinguishing characteristics are logical ones, one special kind of inference would, of course, be the immediate inference of the traditional logic. It is possible to

¹ See, for example, Cohen (1962) and Osgood (1957).

regard the purpose of sections in logic textbooks on immediate inference as being partly to extend the student's understanding of the types of statement involved. In this case a criterion of his understanding of these types of statement would be the ease and certainty with which he makes the appropriate inferences.

At first sight, then, this would be one way of establishing whether indeed PR subjects generally have a poorer grasp of different types of logically important statement than GR subjects. Its chief weakness as an experimental method is that it puts the subject too much on his guard: faced with the question whether 'All A's are B's' implies that all B's are A's, for example, a subject is likely to return the correct negative answer if only because the question would hardly be worth asking otherwise. And even if this difficulty could be overcome - perhaps by asking other questions which seemed even simpler or, as Wilkins did, by including such items in an extended 'syllogism test' - there remains the possibility that a subject might on other occasions make implicitly an inference which he would not make explicitly and after due consideration. Since the processes which are responsible for failures in syllogistic reasoning are likely to be implicit rather than explicit,¹ this seems to be a good reason to prefer an approach in which the subject is able to make the relevant inferences without having his attention drawn clearly to them.

Such an approach was suggested to me by a paper read by P. C. Wason at the Annual Conference of the British Psychological Society in 1968 in which he described in greater detail experiments already referred to in his contribution to Foss (1966). The materials Wason used were sets of cards, each with a letter on one side and a number on the other. Subjects were presented with a statement about each set of cards - laid out with some letters and some numbers face upwards - in which some generalisation about the relationship between the symbols on the two sides of the cards was made, for example, 'If a card has a vowel on one side then it has an even number on the other.' The subject's task was to say which of the cards in the set he needed to turn over in order to establish whether the statement was false.

1 See, for example, the points made in the previous section about the nature of the mistakes PR subjects may be supposed to have made in misjudging the validity of items V6, W11 and W13.

Wason found that at the outset subjects tended to make two kinds of error: (1) they turned over some cards which had no obvious bearing on the question whether the statement was true or not (in the above example, cards with an even number on the exposed side); (2) they failed to turn over some cards which they needed to turn over in order to be certain about the statement's truth-value (in the above example, cards with an odd number on the exposed side). Errors of the second kind tended to drop out with practice but errors of the other kind were remarkably persistent.

Wason was interested primarily in this important and rather puzzling behaviour as an indication of the assumptions subjects made about the verification procedures relevant in such a case. I shall have some comments to make later on about the hypotheses which he offers in an attempt to explain this behaviour. In the meantime I think it is possible to interpret the errors made by Wason's subjects as evidence of an inadequate grasp of the meaning of the kind of statement involved - in the above example, the universal affirmative couched in the hypothetical form preferred by modern logic.

Anyone who has a complete understanding of such a statement ought to be able to carry out certain operations with it, including the rather special kind of operation described by Wason. Amongst other things, he ought to realise (in the above example) that a card with an even number on the exposed side cannot be an exception to the proposed generalisation and so cannot prove the statement false. To fail to appreciate this is to fail to understand that the statement neither says or implies anything about what cards with even numbers have on their other sides. In other words, such a person confuses the statement given with another, stronger statement in which the relationship between having a vowel on one side and having an even number on the other is said to be reciprocal - that is, the statement: 'A card has a vowel on one side if and only if it has an even number on the other'. To put it another way and to bring the potential relevance of this kind of task to the forefront of this discussion, Wason's subjects seem to have been guilty of the assumption that 'All S is P' implies that all P is S. More generally, then, we appear to have in Wason's card-turning task a material which makes it possible for subjects to make illicit inferences at an implicit level, and inferences which can be plausibly construed as reflecting an incomplete understand-

ing of the statement form in question.

I decided, therefore, to use this task to investigate differences in the extent to which PR and GR subjects understand various types of statement involved in Section B of the Valentine test. All the forms of statement are given in Appendix B. The most important, in terms of frequency, are 'All S is P', 'Only S is P' and 'No S is P'. These appear respectively 12, 10 and 7 times in each form of the test - although not always in a straightforward form. The remaining forms of statement occur only once or twice and it would be hard to justify the choice of some of these rather than others on grounds of possible importance in determining the outcome of the test. On grounds of general logical importance, however, two seemed to merit special consideration - those in hypothetical and those in disjunctive form. Given the nature of the task these could not be presented in their simplest, propositional calculus form but were inevitably complicated by having the subject term quantified. It seemed most appropriate to use the universal quantifier and the result was two forms of statement corresponding to those used by Wason in the papers mentioned above and in a more recent paper (1969). Needless to say, this makes it possible to relate my own findings to his.

The five types of statement thus selected were presented in the following form:

- (Ah) 'If a card has a cross on one side, it always has a vowel on the other.'
- (Ad) 'Every card has a capital letter on one side or else an even number on the other.'
- (A) 'All the cards with a square on one side have a Greek letter on the other.'
- (E) 'No cards with a vowel on one side have an odd number on the other.'
- (F) 'Only cards with a small letter on one side have an even number on the other.'

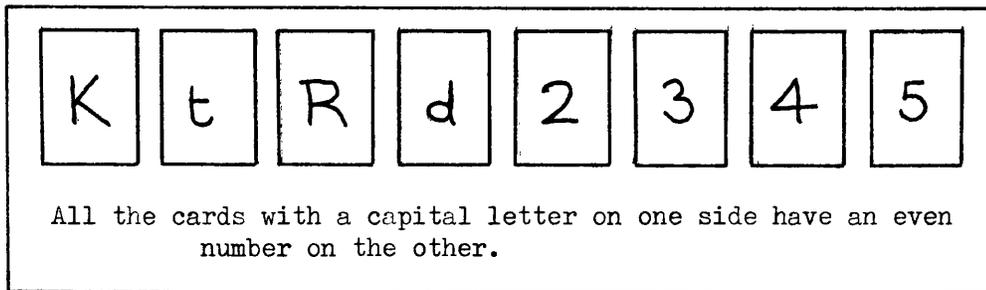
The letters in brackets on the left were the symbols I used to refer to the different statement types. The 'A' and 'E' are, of course, letters used in the traditional formal logic to refer to the universal affirmative and universal negative. The small 'h' and 'd' refer to the hypothetical and disjunctive forms these statements are given in these cases. (As already explained, the Ah form really is simply a universal affirmative in the form preferred in modern logic; the Ad form of statement is essentially different from these two.) 'F' was a letter chosen at random in the absence of a more appropriate symbol - as, for example, a

reversed 'A'. The 'content' of the examples given above is illustrative only: as we shall see, no particular type of content was associated with any particular type of statement.

The next task was to prepare five sets of cards for each type of statement, twenty-five sets in all. In actual fact subjects worked, not with real cards, but with diagrammatic representations of the exposed sides of cards of the kind reproduced, in greatly reduced form, in Figure 2.1 below. Each 'set of cards' was presented on a strip of paper eight inches long and two inches high.

FIGURE 2.1

A 'SET OF CARDS' OF THE KIND USED IN THE 'FIVE TYPES OF STATEMENT' TASK



The purposes to be served by the substitution of diagrams for actual cards were several. It seemed that the purely practical problems involved in presenting twenty-five sets of eight cards to forty-four separate subjects would be very considerable since of course the cards would have to be presented in a certain order and with the correct side uppermost. Not only was this likely to take up a considerable amount of valuable time but rapport with subjects was likely to be lost in the process.¹ Moreover the use of separate strips of paper for each subject appeared to offer an easy and reliable way of keeping a permanent record of subjects' responses: subjects placed a tick below the cards they thought they needed to turn over.

I decided to vary the symbols on the cards to reduce any interference effect between trials, and to vary the number of cards which needed to be turned over for a correct solution in different trials to

¹ A year later a way of coping with the practical problems was devised - under the pressure of the need to use actual cards in order to carry the investigation one stage further.

discourage any tendency towards a mechanical or inflexible approach to the five successive presentations of any one statement type. The twenty-five 'sets of cards' were then arranged in blocks of five, with each statement type being represented in each block and with the order of types of statement within each block being varied in random fashion from subject to subject within each of the two groups (order within a group itself being determined by contingencies such as convenience of time given the other commitments of subject and experimenter.) As an example of the set up, the first subject in each group was presented with sets of cards in the following order: A, Ah, E, F, Ad, E, A, Ad, Ah, F, E, Ad, A, Ah, F, E, Ah, A, F, Ad, Ah, Ad, A, E, F. In this way, of course, I hoped to eliminate any constant errors due to facilitation or interference effects as between one statement type and another.

In addition I prepared two simple sets of actual cards and a 'dummy' strip of diagrammatic cards of the kind subjects were going to have to deal with but with a statement of a type not included in the set of five. These additional materials were designed to help me to explain to subjects the nature of the task and to eliminate any misunderstandings which might arise from the use of representations of cards as opposed to real cards. I also prepared sets of actual cards corresponding to the diagrams on the final five strips in the hope that these might be used to throw some light on the question whether subjects could draw the right conclusions about the truth or falsity of the statements if they were allowed the information they thought they needed - and would get by turning the cards over. I shall say more about this fifth series and the subsequent development of the idea which underlies it later.

It will be apparent to those familiar at least with the earliest stages of Wason's work with the present task that the materials which I have described above differ from those used by him not only in the substitution of diagrams for real cards but also in the use of eight instead of four cards. Originally at least Wason's four cards represented the four possibilities corresponding to the four possible truth combinations of the antecedent and consequent of a hypothetical: $pTqT$, $pTqF$, $pFqT$ and $pFqF$, where 'p' stands for the antecedent, 'q' for the consequent, 'T' for 'true' and 'F' for 'false'. In the statement: 'If a card has a vowel on one side, then it has an even number on the other' the above considerations would call for a set of cards showing, respectively, a vowel, a consonant, an even number and an odd number (with an even number,

an odd number, a consonant and a vowel on the reverse). With a set of cards of this minimal kind the correct choice of cards is always of the same size, not only for the above Ah type of statement but also for the other four types I considered. I have said that I thought it important, or at least useful, to avoid such a situation and using the larger set of cards, where any of the four possibilities can be represented by one, two or three cards in any particular set, enables one to do this. More important, perhaps, the larger set makes it possible for a subject to treat cards of the same kind (all 'pTqF' cards, for example) differently. This is in fact a rather primitive kind of response, at least so far as one can see, and it might seem unlikely that highly intelligent adults, such as my subjects were, would make it. Still, I did not think that one could overlook this possibility and, as it turned out, two subjects, one from each group did turn over less than all of the cards of any given type - I mean regularly and not just on occasion, as might happen purely as a consequence of a perceptual failure. In fact the risk of perceptual failure - failure to notice all the cards of a particular kind - seems to be the only disadvantage of a non-practical kind attaching to the more elaborate set-up which I used. It should be of some interest to use such materials with less able and with younger subjects with a view to establishing at what age or level of intelligence (if any) the primitive kind of response described above commonly appears.

Subjects were tested individually in sessions which usually lasted for about an hour but which in two cases ran on for considerably longer. Only the first part of this period was given over to the task presently under consideration (this part lasting, generally, between 30 and 40 minutes). Any subject who appeared to be tired or distracted or otherwise unlikely to do himself justice was encouraged to give up and return at a later date. (In fact two subjects were lost permanently in this way since neither returned.)

In the course of establishing the necessary degree of rapport between subject and experimenter subjects were told that the aim of the experiment was to identify some of the sources of difficulty which we all experience at times in distinguishing valid from invalid arguments. They were not told of the division of subjects into 'poor reasoners' and 'good reasoners' in case a subject's assumptions about the category into which he fell would affect his responses and so introduce an additional source of uncontrolled variation. Instead, subjects were simply told

that they belonged to a single group chosen with great care to represent all levels of ability in this kind of task - as indicated by performance on Section B of Valentine's test - in the Ordinary Psychology Class as a whole.

Subjects were told that in the past psychologists interested in errors in deductive reasoning had assumed that the difficulty must arise from the failure to perceive relationships between the component parts of an argument. In contrast the experiment in which they were taking part was intended to probe the possibility that the difficulty arose at an earlier, or logically prior, point in the process - in the understanding of the statements which were the component parts. A few words were said about the problems involved in measuring understanding and the experimenter did not claim to be certain that the tasks presented did in fact measure this. It was suggested, however, that these tasks were the best means at present available for this purpose. Finally, by way of preamble to the explanation of the task itself, the relationship between the five types of statement to be presented and the Valentine test's Section B was explained: that is to say, subjects were told that these were the five most important types of statement involved in that test.

The explanation of the task proper was begun by showing the subject the first of the simple sets of cards - a set of three cards with the numbers 2, 4 and 5 respectively on their exposed sides and a strip of card on which was printed the sentence: 'The cards with an even number on one side have a vowel on the other'. The subject was asked to say what he would do in order to discover whether the sentence was true or not as applied to the set of three cards in front of him. The simple answer looked for was: 'Turn the cards over' or, of course; 'Turn the cards with even numbers over'. To my surprise some subjects proposed an incorrect procedure even in this very simple case (for example, they suggested turning the card with the odd number over) but the mistake was not pointed out to them as a mistake; instead the experimenter said something non-committal such as: 'Yes, you would turn some or all of them over' - so long as the elementary idea of turning the cards over seemed to have been grasped.

Next the subject was presented with a set of four cards bearing the following symbols on their upturned faces and in the following order:

α , t, +, \square . The accompanying sentence read: 'Every card with a Greek letter on one side has a square on the other'.

This set of cards carried a great deal of the weight in the business of explaining the terms of the problems the subject would be asked to solve. One or two sources of irrelevant difficulty were cleared out of the way for a start. The subject's attention was drawn, for example, to a display card on which all the Greek letters to be used in the entire series of cards were written. (These were α , β , δ , and π .) He was told that, if in doubt, he was to regard any other letter as not-Greek or Roman. It was emphasised that difficulties arising from typography, as it were, were unintentional and irrelevant and that the subject should always remove any uncertainties arising from this source at once by asking what a particular symbol was supposed to be. (As it turned out, difficulties of this kind hardly ever occurred.)

The main burden of the explanation was as follows. The sets of cards the subject would have to deal with would normally have eight cards instead of four (as in the set in front of him) but they would be divided, as the present set was, into two equal groups laid out to left and right. The cards in the left hand group would be showing their 'face', as it were, and the cards on the right their reverse. The importance of this was that the subject knew what kind of symbols were on the unseen sides of the cards in either group: they were the kind of symbol appearing on the exposed sides of the other group. Heavy emphasis was placed on the fact that it was not possible to infer exactly what would be on the other side of any particular card: in the four-card set in front of him one could infer that the two cards on the left had geometrical figures on their unseen sides but not, for example, that one would have a cross and the other a square. Similarly, the cards on the right would have letters on their unseen sides but one or both or neither of them might be Greek letters. It was also emphasised that no one-to-one correspondence between 'face' and 'reverse' could be assumed - so that, for example, of two cards with even numbers on the exposed sides one might have a vowel and the other a consonant on the unseen side.

The subject was then asked to decide which of the cards in the set of four he needed to turn over to discover whether the statement presented was true or not of that set of cards. Once again, no clear indic-

ation was given as to the correctness or otherwise of his selection, provided, of course, as was always in fact the case, that the idea of turning cards over had been grasped.

Finally, the subject was shown the 'dummy' strip of cards in diagrammatic form and the terms of the problem were run over again. The subject was told that the statement shown on this strip was not one of the types he would be asked to deal with but in fact a more difficult one. He would be asked to imagine that these were cards with symbols on the back and to indicate which ones he needed to turn over in order to discover the truth or falsity of the statement by placing a tick underneath them. He was reminded of the division of the set into two half-sets of four and of the inferences he might, and might not, draw from this about the symbols on the unseen sides. It was pointed out that the number he needed to turn over might in principle be anything from nought to eight but that the experimenter had tried to arrange that the number would vary in an unpredictable way from set to set, so that no assumptions could legitimately be made as to the number required for any particular type of statement. He was also to try to cover all possibilities at one attempt - so that strictly speaking the question was: 'Which cards would it be relevant to turn over? Which cards might affect the issue? (This to meet the natural objection that with a set of real cards one would turn them over one at a time and could stop with the first one which proved the statement false.)

The subject was told that his responses would be timed but that the object of this was to compare, not his times with those of other individuals, but the average times, for the group as a whole, on the five types of statement. Consequently, it would suit my purpose best if he took enough time to ensure that his response was correct but no more. Providing everyone applied this principle consistently, differences between individuals would be of no account. Finally, in this connexion, I described an 'emergency' procedure which had suggested itself during preliminary trials with the material. Sometimes, it was suggested, due to fatigue or distraction, the subject might find himself completely unable to decide, on rational grounds, which cards needed to be turned over. In the event of a 'mental block' of this kind, the subject was to bring that trial to an end by placing a tick under all eight cards and so avoid a latency of response which bore no relation to the difficulty of the task. (In the event this 'emergency procedure' was used extremely rarely

- as, of course, was anticipated.)

In the experiment itself subjects were not told whether they were getting the right solution to the various problems, but signs of anxiety were met with reassuring noises on the part of the experimenter - who would remark, for example, that everyone experienced some degree of difficulty with this kind of task. Because of the absence of 'feedback' about the subject's success or failure little improvement was expected over the five trials (although, as we shall see, some did in fact occur). The point of the fivefold replication was said to be - as it was in fact - the provision of a relatively stable estimate of the relative difficulties of the different types of statement.

At the end of the fourth series of five sets subjects were told that the final series would involve an additional element. After they had dealt with a set of diagrammatic cards, a set of actual cards bearing the same symbols on their exposed sides would be presented. Subjects would then be asked to turn over the cards they had ticked on the corresponding strip of paper and say whether the statement was true or not.

In the event, some subjects found themselves wanting to turn over cards they had not ticked, or else no longer convinced that they needed to turn over all the cards they had ticked. To avoid any loss of rapport, and because this seemed a useful source of additional information, I permitted subjects in the first of these situations to turn over additional cards, requiring them, however, to signal the fact by placing a tick with a plus sign in front of it under the appropriate diagram or diagrams on the test strip.

Finally all subjects were asked to indicate their conclusions about the truth or falsity of the statements in the fifth series by writing 'true' or 'false' on the appropriate test strip.

One final point about the general conditions of the experiment should perhaps be made. Aware of the dangers of distorting the outcome in the expected direction¹ I deliberately tried to avoid getting to know which subjects belonged to which group - to such an extent that even in the second phase of the experiment I was still unable to guess with any degree of certainty which were GR subjects and which PR - at least at the outset!

1 See, for example, Friedmann (1967).

CHAPTER THREE

THE 'FIVE TYPES OF STATEMENT' TASK, PART TWO:

RESULTS, ANALYSIS AND DISCUSSION

Summary In view of the large numbers of errors made by members of both groups, the time taken to complete the card-turning task is presented as a measure of the perceived, as opposed to the actual, difficulty of the different types of statement. There were no significant differences between the groups in this respect. Significant differences were found, however, between the statement-types within each group: in the PR group the Ad and F types of statement were significantly more difficult than the other three, while the same was true of the GR group except that the difference between Ad and Ah was not significant. The question whether the apparent discrepancy between the perceived and actual difficulty of the Ad statement-type might be due to the ambiguity of the disjunctive is discussed.

In section 3.2 the performance of the two groups is compared in terms of right and wrong responses, four methods of assessment of different degrees of specificity - and, it is argued, validity - being employed. The outcome of this central part of the analysis, though not entirely unequivocal, suggests that the PR group, as predicted, understand the A and F types of statement less well than the GR group. The view that the misunderstanding involved takes the predicted form of supposing that the relationship between the 'terms' of these statement-types is reciprocal rather than one-way is strengthened by the discovery of significant differences between the groups in the success with which they deal with the right hand half-set of Ah and the left hand half-set of F. Attempts to find additional support for this conclusion in terms of the kinds of error made in the relevant half-sets and in the extent to which the two groups treated A and F statements as if they were equivalent both failed to reveal significant differences.

In section 3.3 the two groups are shown to be extremely similar in the extent to which their responses varied over the five trials. The PR group is shown to make 'matching responses' significantly more often than the GR group in the four statement-types other than E (where the correct response is identical with a matching response). Different interpretations of this result are considered.

In section 3.4 evidence is presented bearing on the assumption, on which the present experiment was based, that the 'five types of statement task' is a measure of the extent to which subjects understand the statement-types involved. Drawing on the results of Wason and Johnson-Laird, as well as those obtained in the present study, it is suggested that the assumption is a dubious one when comparisons are made between statement-types. Fortunately, there appears to be less difficulty in accepting the assumption when it is used as a basis for comparisons only within statement-types. Finally, in this section, a comparison is made between my results and those of Wason and Johnson-Laird in terms of pattern of response and location and type of error. A very large measure

of agreement is revealed. A hypothesis is developed to account for certain peculiarities in the distribution of errors and types of error, and the 'over-determination' of responses which this new hypothesis appears to imply is taken to suggest the need for some other method of determining which of the various possible agencies is actually responsible for producing the results obtained. Such a method, it is suggested, is offered by the task included in the fifth series of the present experiment and elaborated in the experiment described in the following chapter.

In section 3.5 the additional features embodied in the fifth series of trials are described and some results presented which suggest that the misinterpretation of the F type of statement as implying the corresponding A statement may indeed be important in producing the errors in the selection of cards discussed in previous sections. An unexpected but, it is suggested, unimportant confusion seems to have been responsible for errors in the Ad case. There also appear to be tentative grounds, from the Ad as well as the A case, for suspecting the operation, in the verification task, of a directional effect similar, at a perceptual level, to the effect described by Wason (1969). Finally, some evidence is presented bearing on the question whether the nature of the task is affected when real cards are substituted for the diagrammatic ones used in the main part of the experiment. Various points are made about the potentialities of this fifth series task as a device for answering some of the questions posed by the results obtained elsewhere in the experiment.

3.1 Differences in the time taken to complete the task Time taken was one of the two variables in terms of which it was hoped to detect differences between the GR and PR groups. It was difficult to be sure in advance whether this or the error score would be the better measure, or whether, indeed, some index incorporating both might have to be devised. It is true that the work reported by Wason (1966) would have suggested the error score, but Wason had at that time used the material with only one type of statement - the Ah one - and it seemed wise to cover all eventualities. As it happened, few subjects achieved a level of success in which so few errors were made that the time taken might have been deemed to be the appropriate measure of difficulty.

The stop-watch was started as the paper strip bearing the card diagrams was placed in front of the subject and stopped when the subject indicated that he was finished - by saying something such as 'All right' or simply raising his head, as previously arranged. It would be idle to pretend that the accuracy of measurement obtained in this way was of the order suggested by the tenth parts of a second in terms of which the time taken was recorded. On the other hand, the error would have been a relatively constant one and unlikely to be of great importance for two reasons. In the first place, the times involved were relatively long ones and the error proportionately small: Table 3.1 below shows that the average

time for both groups of subjects over all five types of statement was in excess of 20 seconds. In the second place, our concern is with the relative, and not with the absolute, times taken to complete the task with different kinds of statement, so that an error which is likely to be the same for all types of statement and for both groups is of no practical consequence. And finally, the general adequacy of the time-measurement seems to be confirmed by the mean times for all types of statement taken together: in both groups these decrease progressively from one trial to the next.

I have said that the number of errors made by my subjects was generally too great to justify the use of time as the means of comparing the difficulty presented for members of the two groups by the different types of statement. It is indeed obvious enough that a subject who completes a task quickly but not correctly cannot seriously be said to have had less difficulty with the task than another subject who takes longer but makes fewer errors. On the other hand, it seems reasonable to suggest that the time taken to complete a variety of tasks, none of which is completed without error, and all of which involve essentially the same components, reflects the subject's perception of the difficulty of the various tasks: he will take longer to complete those tasks which seem to him the more difficult or about which he feels less confident. And inasmuch as this seems of some interest in an area about which as yet so little is known, I present the mean times taken for the five types of statement on successive trials and as a whole in Table 3.1 below.

TABLE 3.1

AVERAGE TIMES (IN SECONDS) FOR THE FIVE TYPES OF STATEMENT, SEPARATELY AND TOGETHER, ON FIVE SUBSEQUENT TRIALS

	Trials	1st	2nd	3rd	4th	5th	All
	Ah	24.3	21.9	22.7	17.4	16.8	20.6
	Ad	28.2	28.7	23.2	18.2	17.5	23.3
GR group	A	21.6	23.0	18.3	17.4	15.7	19.2
	E	23.7	22.8	15.6	17.8	17.0	19.4
	F	35.5	25.7	23.7	20.4	23.4	25.7
	All	26.7	24.4	20.7	18.2	18.1	21.6

Continued overleaf

	Trials	1st	2nd	3rd	4th	5th	All
PR group	Ah	29.6	18.4	21.0	16.6	18.0	20.7
	Ad	36.8	23.4	25.2	22.6	24.4	26.4
	A	24.5	21.0	18.8	21.6	16.1	20.4
	E	23.7	22.5	20.6	17.9	17.5	20.4
	F	29.5	25.1	22.3	25.7	23.6	25.2
	All	28.8	22.1	21.5	20.8	19.9	22.6

It will be apparent that in both groups the statements which took longest were the Ad and F types, and we may conclude that these were perceived by the subjects as presenting most difficulty. Non-parametric tests of the significance of the differences between the various statement types were made and the relevant p values are given in Table 3.2 below. My procedure was to apply the simple, and not very powerful, Sign Test (Siegel, 1959) in the first instance and to go on to apply the Wilcoxon Signed-Ranks Test, with its greater power, only in those cases where the result with the Sign Test was not significant. Inter-group differences were tested in the same way: none were significant.

TABLE 3.2

P VALUES FOR DIFFERENCES IN MEAN TIMES TAKEN WITH DIFFERENT TYPES OF STATEMENT

	<u>GR group</u>				<u>PR group</u>			
	Ad	A	E	F	Ad	A	E	F
Ah	N.S.	N.S.	N.S.	p=.004	p<.002	N.S.	N.S.	p<.02*
Ad		p=.016	p=.016	N.S.		p<.002	p<.002	N.S.
A			N.S.	p<.002			N.S.	p<.01*
E				p<.002				p<.02*

*Wilcoxon Signed-Ranks Test: all other p values on Sign Test

It is hardly surprising that the F type of statement should appear to subjects to be relatively difficult since, as we shall see, it was the type of statement which produced most errors and was, in that sense, actually most difficult. The case is otherwise with the Ad type of statement which proved to be, in terms of error, less difficult than either Ah or A. Particularly striking, perhaps, is the very long mean

time for the first Ad trial taken by members of the PR group. It would appear that the disjunctive type of statement struck these subjects as of quite exceptional difficulty, at least at first sight.

It is perhaps possible to explain the special position occupied by this type of statement, however, in terms of its ambiguity. Logicians customarily distinguish two senses of 'or', the inclusive sense, in which the truth of the statement that an object has one quality or another is consistent with its having both, and the exclusive sense, in which this possibility is excluded. Subjects in the experiment now under discussion were not told in advance which sense of the disjunctive they were expected to assume - if only because they were not told this in the Valentine test either. The long initial time, especially in the case of the PR group, may reflect this uncertainty and the attempts of at least some subjects to reach a decision on the point. In later trials the question may have continued to make itself felt, either consciously or otherwise.

It is not possible to be sure whether this is the true explanation of the discrepancy between the time taken by both groups on disjunctives and their success with them as measured in terms of errors - between, as I have suggested the real and perceived difficulty of this type of statement. Only two subjects actually voiced a doubt about the sense in which they were to take the words 'either...or else' (and in these cases, incidentally, I gave no clear guidance). Subjects to whom the question did explicitly occur may have answered it either in terms of the naturalness of the two possibilities (the inclusive sense being, I think, the more natural) or else in terms of the experimental task. The second alternative would have called for a fairly elaborate chain of reasoning for it involves (1) the recognition that if the disjunctive is interpreted in the exclusive sense, all eight cards in a set must be turned over, and (2) the belief that such an outcome is inconsistent with the terms of the problem as set out in the instructions (the number of cards requiring to be turned over was to vary unpredictably from trial to trial for any particular type of statement) or perhaps with the arrangements likely to have been made by the experimenter.

Whatever the explanation the great majority of subjects do seem to have come to the conclusion that the appropriate sense of the disjunctive was the inclusive one - as we shall see when we consider the patterns

of choices in detail and subjects' responses in the verification part of the fifth series. The same was true twelve to eighteen months later when the same subjects did a similar task in conditions which encouraged them to ask questions and express doubts. Wason too has subsequently reported results which confirm that most subjects understand the disjunctive in the inclusive sense: only 3 out of his group of 10 said, when asked, that the possibility of interpreting the disjunctive in an exclusive sense had occurred to them (Wason, 1969). It is of great interest to note, too, that Wason's subjects seemed to be confident but frequently wrong about the Ah type of statement (the universal affirmative in hypothetical form) and less confident but more often right about the disjunctive. This seems to be the clearest independent confirmation of the result discussed above.

3.2 Difficulty measured in terms of errors Before we consider the problems involved in establishing an error score as a measure of the difficulty of the various types of statement, it might be thought appropriate to say what the correct response in each of the five types of statement was, and to explain why the designated response was the correct one. In this way the nature of the task and the different possible ways of scoring responses in terms of adequacy may become apparent.

The Ah type of statement, as we have seen, took the following form: 'If a card has a vowel on one side then it always has a circle on the other.' The subject in this example would have four 'cards' on his left with letters on the exposed side (at least one with a vowel) and he is asked to assume that these cards have geometrical figures on their unseen sides; on his right he has four cards with geometrical figures on their exposed sides (at least one being a circle) and of course he is asked to assume that these cards have letters on their other sides. The statement will be false if and only if there is at least one card with a vowel on one side and a geometrical figure other than a circle on the other. The only cards which could be of this kind are those on the left which have a vowel and those on the right which do not have a circle. Accordingly, it is only these two kinds of cards which it is relevant to turn over - and in order to be sure that he does not miss a card of the critical kind he must turn over all the cards of these two types.

If he fails to turn over a card of one of these two types (or rather - as the problem was presented - if he fails to place a tick under

a diagrammatic card of either of these types) he makes what I came to call an 'error of omission' (or 'EO' for short). If he turns over a card which is not of either of these two types, he makes an 'error of commission' (or 'EC'). Wason (1966), working with statements of the Ah type found errors of commission common and very persistent. On the other hand, as I shall suggest later, they are the less serious of the two types, at least in terms of their implications for success in a situation in which subjects are set actually to detect statements which are false. For my research, on the other hand, they have a special significance in the Ah and A types of statement (the two forms of the universal affirmative) because they may be thought to reflect a subject's mistaken assumption that the relationship said to hold between the two classes of cards in a universal affirmative is a reciprocal one. In other words, it may be supposed to reflect the belief, referred to in previous chapters, that 'all S is P' implies that all P is S or, in terms of the hypothetical, that 'if a thing is an X it is also a Y' implies that if a thing is a Y it is an X. To be precise, this seems to be a possible interpretation of EC's in the right-hand half set.

Disjunctive statements (Ad) were represented by the following example (amongst others): 'Every card has a cross on one side or else a vowel on the other'. We have already remarked on the possible ambiguity of this type of statement and on the fact that if it is interpreted in the exclusive sense, a correct response would call for the turning of all eight cards. This is because the statement in the exclusive sense will be false if and only if (1) there is at least one card which has neither a cross on one side nor a vowel on the other, or (2) there is at least one card which has a cross on one side and a vowel on the other. The first of these conditions requires one to turn over all cards with a geometrical figure other than a cross on the exposed side (to check that they do have vowels on the other side) and all cards with a consonant on the exposed side (to check that they do have crosses on the other side). The second condition requires one to turn over all cards with a cross on the exposed side (to check that they do not have a vowel on the other side) and all cards with a vowel on the exposed side (to check that they do not have a cross on the reverse side). In other words, the two conditions together oblige one to turn over all eight cards. If the disjunctive is interpreted in the inclusive sense, on the other hand, only the first of the above two conditions is relevant - so that the correct response, in the above example, would be to turn over all cards with

geometrical figures other than crosses and all cards with letters other than vowels (i.e., of course, consonants). In scoring responses to this type of statement I took the inclusive response to be the correct one for reasons already mentioned: that the inclusive sense seems the more natural¹, that it makes better sense in the present experimental situation, and that, presumably for one or other of these reasons, few subjects, if any, seem to have interpreted it in the exclusive sense. (In each group all eight cards were marked on some twenty occasions out of the total of 110 occasions on which such a response might have been made to a disjunctive. In no case, however, did a subject say that the disjunctive of the fifth series was false although one of the actual cards used in this series did combine the two characters mentioned.)

The universal affirmative in its plain form (A) may be represented as follows: 'All the cards with a vowel on one side have an even number on the other'. As this is simply another form of a statement which may be represented by the Ah type already discussed, the type of response which was appropriate in that case is also appropriate in this: that is to say, in terms of the present example, the cards to be turned are all those with a vowel and all those with an odd number on their exposed sides.

The universal negative (E) may be represented by the following statement: 'No cards with a cross on one side have a Greek letter on the other'. The kind of card which would make this statement false is clearly one which ^{does} have a cross on one side and a Greek letter on the other, and the subject must therefore indicate that he would turn over all cards with a cross and all those with a Greek letter on their exposed sides since all of these (and only these) may prove to be cards of the kind which would falsify the statement.

Finally, statements of the F type may be represented for just

1 There appears to be something very unempirical about the bald statement that the inclusive sense of the disjunctive 'seems the more natural'. At least one needs to add the qualification that it seems the natural one in the present context. (On the role of context see Carney and Sheer, 1965). Obviously there are devices which we ordinarily employ when the context leaves the issue in doubt and we wish to avoid such a state of affairs. Such, for example, would be the use of the phrases, 'or both' and, 'but not both' to indicate the inclusive and exclusive senses respectively.

now by the statement: 'Only cards with an even number on one side have a vowel on the other'. This will be false if and only if there is at least one card which has an odd number on one side and a vowel on the other. The correct response consists, therefore, in turning over cards with characters of either of these kinds on their exposed sides, since the former might have a vowel on the other side and the latter an odd number. As the F type of statement is the 'mirror-image' (not, as we have seen, the converse) of the universal affirmative, it is to be expected that subjects will have difficulties with it which are the counterpart of those found by Wason in the case of Ah type statements. And just as it seemed plausible to interpret errors of commission in the right-hand half set in A and Ah cases as involving a confusion about the meaning of the universal affirmative, so, in F type statements, we may think it appropriate to interpret errors of commission in the left-hand half set as reflecting a confusion about the meaning of these statements - and, in particular, a mistaken assumption that a statement of the form 'Only S is P' means or implies that all S is P. If anything, we might expect this confusion to be even more prevalent, not only because there are situations in which (as Black, 1946, p. 115 remarks) it is reasonable, on empirical grounds, to assume that 'only' means 'all' as well, but also because there are closely related forms of statement in which the universal affirmative is part or all of the meaning. Thus 'Only the X's are Y's' implies that all the X's are Y's (as well as that all the Y's are X's, the change of meaning being effected by the insertion of the definite article between 'only' and 'X's'); and 'The only X's are Y's' (with the definite article preceding 'only') is equivalent in meaning to 'All X's are Y's'.

If, in referring to such a statement as the one at the top of this page, we say that cards with even numbers on their exposed sides are 'cards with the characters first named', and that cards with vowels on their exposed sides are 'cards with the characters named second' - the two categories together constituting, of course, the class of cards with named characters - it is possible to make certain rather general remarks about responses to the tasks now being described and, in particular, to summarise the properties of a correct response to the five types of statement as follows (availing ourselves throughout of the special form of expression referred to above, 'only the X's', to mean all X's and only X's):

<u>Statement Type</u>	<u>Left hand half set</u>	<u>Right hand half set</u>
Ah and A	Turn only the cards with first named characters	Turn only the cards without named characters
Ad	Turn only the cards without named characters	
E	Turn only the cards with named characters	
F	Turn only the cards without named characters	Turn only the cards with the characters named second

Differences in the success of the two groups assessed by different methods

There is no self-evidently correct way of assessing the respective degrees of success with which the two groups tackled the task. Reflection suggested four different methods of varying degrees of adequacy, all of them providing interesting data about the performances of the groups. These will now be described in an order ranging from the most 'global' (and least adequate) to the most detailed (and, perhaps, most adequate).

In the first method account is taken only of the numbers of subjects in each group who made the correct selection of cards to turn over on all five occasions for any one statement type. It is only such subjects, it might be said, whose understanding of the relevant statement is complete and firm. Table 3.3 presents the relevant figures.

NUMBER OF SUBJECTS IN EACH GROUP MAKING THE CORRECT RESPONSE IN ALL FIVE TRIALS FOR DIFFERENT STATEMENT TYPES

	<u>Statement Type</u>				
	Ah	Ad	A	E	F
GR group	-	8	-	4	1
PR group	1	4	1	3	-

This first method of assessing the performances of the two groups is the only one which gives the PR group the advantage in any detail over the GR group. This in turn is due to the prominence it gives to the performance of one of the PR subjects who succeeded in making the correct response to three of the five types of statement. This was in fact the best performance of any subject from either group, no other subject achieving success on more than two statement types. (To be specific, success on all five trials in both Ad and E was achieved by two GR subjects and one PR subject, in both A and F by one GR subject,

all other successes being 'singletons').

The success of the PR subject first mentioned above is so much in contradiction to the expected (and, by other methods, established) superiority of the GR group that it seemed worthwhile to check that she had not been included in the PR group by mistake or, more plausibly, that there was not some reason to suspect that her performance on the Valentine was attributable to the operation of some purely temporary factor. In the absence of any explanation of either of these kinds for the discrepancy, her exceptionally good performance on the 'five types of statement' task must be taken to represent strong evidence for the operation, in a deductive reasoning task, of factors quite different from those involved in the experimental task.

Reference to Appendix E will show that this subject (LD) had an AAA score of 65 and a Valentine Section B score of 47. As these are T-scale scores, the gap between them is equal almost to two standard deviations and as such almost twice as large as the minimum I had set in selecting subjects for my PR group. On the other hand, judged in absolute terms, hers was clearly not one of the poorest performances on the Valentine test, Section B. It is, of course, just below the average for the class as a whole, and it is just possible that there are reasons for believing that it would not have fallen below the average if rather temporary factors had not played a role in her performance.

In her answer to item 47 in the questionnaire which all subjects were asked to complete (as described previously in this thesis) she said that she was 'physically below par' and 'emotionally upset' on the occasion when she did Form V of the Valentine. In answer to item 44 ('If you found one Form of the Valentine test easier, can you say why this was so?') she wrote: 'In Form V Section B the reasons for certain conclusions being illogical did not appear to be complete; however, I had just had a rather unsettling afternoon and was very tired, so perhaps merely inability to concentrate and reduce the problem to its basic factors was the reason'. This looks, perhaps, like 'prima facie' grounds for supposing that her performance on Form V was unrepresentative. On the other hand, however, she actually did better on Form V than on Form W, and since this is entirely consistent with the performance of the group as a whole (she did Form V after Form W) it did not seem to me at the time when I was forming my experimental groups, and it does not seem to me now, that

I should be justified in excluding her from the PR group. Needless to say, in the light of her performance on the task now being described I was very much looking forward to further investigation of her case and especially disappointed, therefore, that she should turn out to be the only subject unable to return for the second phase of my research twelve months later. By that time she was married and living in Belfast.

It will be clear that, however interesting it is to know just how many subjects managed to achieve an error-free performance on all five trials on any one statement-type, a comparison of the two groups on this basis is scarcely a satisfactory procedure. It is not only that, as we have just seen, the performance of exceptional individuals is likely to distort the overall picture but more generally that the error component in the method of assessment is likely to be unduly large. In the first place, the subject who takes a little time to adjust to the experimental situation or to 'realise' what the task requires is likely to make errors on the first trial and to find himself grouped, on this first method of assessment, with other subjects who make errors on all five trials. Secondly, the five types of statement task, at least in the specific form given it in the present research, clearly involves a perceptual component: subjects have to perceive and recognise the various cards as belonging to one category or another, and although, of course, the task would be an unsatisfactory one for my purposes if the perceptual task was a difficult one, it seems clear to me that errors (especially errors of omission) were occasionally to be attributed to a failure at a perceptual level. (Some subjects had a single error of omission in the third or fourth trial in an otherwise faultless series.) And although these are less easily identified, it seems quite likely that a few successes are to be explained in terms of errors of commission which were not made because the subject did not perceive or recognise a card which he would otherwise have (mistakenly) turned over.

Some of the above difficulties would be reduced, if not removed altogether, if we considered the performances of subjects on different trials separately, instead of their performances on all five trials taken together. In particular, it might be suggested that success on the fifth trial should be regarded as the best criterion of success on the task: by then, of course, subjects would have had time to adjust to the experimental situation and to come to terms with the task, and the incidence, and effect, of errors with a purely perceptual source would be unlikely

to be very large.

There appears to me, however, to be one crucial objection to this particular approach and it is that it conspicuously fails to utilise all the available evidence about the degree of confidence or certainty with which a subject appears to have grasped the meaning of a statement. Two kinds of case come to mind: a subject who has made the correct choices in four cases out of five may plausibly be said to have a better understanding of the statement in question than a subject whose eventual success is preceded by four failures; and a subject who achieves success on the third and/or fourth trial but not on the fifth may be thought, at least, to have had an insecure grasp of the meaning of the statement and its implications for the experimental task when a subject who fails on all five trials has, so far as one can tell, no grasp at all. In Table 3.4, therefore, I present the number of correct responses made by each group as a whole, each trial being considered separately. If it is correct to regard the number of trials on which any one subject makes the appropriate response as an index of the certainty with which he perceives the relationship between the statement and the task, then the same ought to be true of each group as a whole.

TABLE 3.4

TOTAL NUMBER OF CORRECT RESPONSES FOR EACH GROUP FOR
EACH TYPE OF STATEMENT

	<u>Statement Type</u>				
	Ah	Ad	A	E	F
GR group	14	66	15	72	22
PR group	12	44	9	53	10

The differences between the groups, as revealed on this second method of assessment, are in the expected direction but clearly still not large enough to be significant. Perhaps the feature of these results which calls most for comment is the relative ease with which subjects appear to have tackled the disjunctive and universal negative types of statement. The latter in particular seems to call for comment, but I defer this till I come to comment more generally on the performances of the two groups on this task. In the meantime, it will perhaps be apparent that this second method of assessment, although undoubtedly

an advance on the first, is still not the best available. It does take account of the varying degrees of certainty with which subjects handle the five types of statement, but it still does not utilise all the available information. In particular it takes no account of the success with which subjects deal with the two half-sets of cards or of the extent to which they make errors of both kinds (EO's and EC's) or of only one kind. (It seems reasonable to suggest that a subject who deals correctly with one half-set does better than a subject who gets both wrong, and a subject who makes only one kind of error does better than a subject who makes errors of both kinds.)

To take account of these aspects of the problem situation I developed a scoring system in terms of which a subject could score a maximum of four points for each trial - therefore a maximum of twenty points for a complete series of five trials. From the four points awarded for a correct response on any one trial one point was deducted for each type of error in each half-set. This means, for example, that a subject who made one or more errors of omission or one or more errors of commission in only one half-set would have a score of three for that trial. A subject who made both types of error on one half-set, or one type of error in both half-sets, or one type of error in one half and the other type of error in the other would have a score of two. A score of one would be awarded to a subject who made both types of error in one half and one in the other, and, finally, of course, a subject who made both types of error in both half-sets would score nothing. Table 3.5 presents the average scores of the two groups on the five types of statement using this third method of assessment.

TABLE 3.5

AVERAGE SCORES OF THE TWO GROUPS ON THE FIVE TYPES OF STATEMENT TAKING ALL FIVE TRIALS TOGETHER (MAXIMUM POSSIBLE SCORE FOR EACH TYPE = 20)

	<u>Statement Type</u>				
	Ah	Ad	A	E	F
GR group	13.3	14.7	13.7	17.0	12.4
PR group	12.0	12.0	11.7	16.1	9.9
Difference	1.3	2.7	2.0	0.9	2.5
Wilcoxon T	84.5	79.5	53.5	76.5	60.5
N Ranks	21	21	22	21	21
p (two-tailed)	N.S.	N.S.	<.02	N.S.	N.S.

It will be seen that the statistical test used in deciding whether any of the differences between the groups was too large to be attributed to the operation of chance factors was the Wilcoxon Signed-Ranks test (Siegel, 1956, Hays, 1963) and that the p-value quoted is for a two-tailed test. It seemed clear that the appropriate statistical test for the data now being considered was a non-parametric one, for it is extremely difficult to justify any particular assumption about the nature of the distribution of scores in the population from which the present samples are drawn. In view of the experimental design, the appropriate test appeared undoubtedly to be the Wilcoxon, a test, according to Hays (*ibid.*, p. 635) of 'very high power-efficiency compared to other methods designed specifically for the matched-pair situation'. The T and N Rank values were calculated in the way described by Hays rather than that apparently favoured by Siegel. That is to say, zero differences were not dropped from the analysis but were given the average (lowest) rank, half the sum of their rank values then being assigned to the smaller set of differences (on which the T value is based). Where N ranks in the above table is 21 instead of the expected 22, this is because the number of the zero differences was odd and, in accordance with the procedure described by Hays, one was discarded. The p values were read from the Wilcoxon table reproduced in Siegel and, in extended form, in Guilford (1965).

As to the question whether a one-tailed or a two-tailed test is the appropriate one in the present research, it is extremely difficult to be certain whether the experimenter's expectation that the PR group would do less well on at least some of the five types of statement than the GR group amounted to a prediction that this would be the case. Certainly, it would have been a major surprise if any significant differences had been in the opposite direction. The opinions of the authors of statistical texts seem to vary in their interpretations of the conditions which justify the use of a one-tailed test from the rather strict (Edwards, 1968) through the intermediate (Diamond, 1959, Guilford, 1965, Hays, 1963) to the rather lenient (Siegel, 1956, McNemar, 1959). In view of this state of uncertainty the fact that my prediction - if such it could be called - would have been of the form, 'If there are any differences between the groups on the five types of statement, they will favour the GR group' I decided that a two-tailed test would be the appropriate one.

In this and in subsequent Tables I have noted p values equal to

or less than .05, all others being regarded, of course, as not significant. The justification for considering differences with a p value as large as this to be significant is partly that it is in line with an established convention in psychological research. More important, perhaps, any leniency implied in this decision tends, so far as the power of the test is concerned, in a direction opposite to the one implied by the decision to employ a non-parametric test and to consider both tails of the sampling distribution. And finally the need to take precautions against the possibility of a Type I error does not seem to be a particularly pressing one in a study which is admittedly exploratory rather than definitive in character. (Compare Diamond, 1959, p. 117) The intention was to follow up any differences which seemed to be significant rather than to accept them as facts complete in themselves.

Having said all this, the important point to be observed in the results reported in Table 3.5 is obviously that the difference between the groups in the success with which they tackled the A type of statement was large enough to be significant and that no other difference was. In particular it is a rather surprising fact that the other form of the universal affirmative - Ah, the hypothetical form - is so far from giving a significant difference that its T value is larger than for any other type of statement. (An advantage of the procedure for calculating T described by Hays is that it facilitates comparisons of this kind by reducing to two the number of values taken by N ranks.)

It remains to describe the fourth and most specific method of assessing the performances of the two groups on the five types of statement. In this method account is taken not simply of the types of error made but of the number of errors of both types. The difference between the two approaches can be described by saying that whereas the third method focuses on the four different operations which are the components of a correct response (see the summary on p. 70 above) and deducts a point for each operation not carried out correctly, the fourth method focuses on the treatment of individual cards, asking of each whether it was dealt with in the appropriate manner or not, and counting the number of cases in which it was not (in which it was turned over when it should not have been, or not turned over when it should have been). It will perhaps be apparent that the results obtained by these two methods will be different only if subjects do not always treat all the cards of a particular type in the same way, and, as we shall see, the question whether one of these

methods is superior to the other depends on the interpretation we put on cases where all cards of a particular type are not treated in the same way. In the meantime, Table 3.6 presents the average number of errors per statement-type for each group.

TABLE 3.6

MEAN NUMBER OF ERRORS OF THE TWO GROUPS ON THE FIVE TYPES OF STATEMENT TAKING ALL FIVE TRIALS TOGETHER. (MAX. POSSIBLE FOR EACH TYPE = 40)

	<u>Statement Type</u>				
	Ah	Ad	A	E	F
GR group	11.9	9.1	11.9	4.6	13.9
PR group	14.5	14.1	14.9	6.7	18.8
Difference	2.6	5.0	3.0	2.1	4.9
Wilcoxon T	71	80.5	58.5	85	53.5
N ranks	21	21	21	22	22
p (two-tailed)	N.S.	N.S.	<.05	N.S.	<.02

The difference between the two groups on the A type of statement remains significant on the fourth method of assessment, though with $p < .05$. At the same time, the difference on the F type of statement becomes significant for the first time. This makes the outcomes of the two methods seem to be more discrepant than perhaps they are, for the difference on F using the third method falls only just short of significance at the $p < .05$ level. (T = 60.5 when the critical value for N ranks = 21 is 59.)

Such discrepancy as exists is, as we have seen, to be attributed to the fact that in the fourth method of assessment different weights are given to cases in which all the cards of a particular kind are treated in the same way (turned or not turned) and cases where some are turned and some are not. If there were no cases of the second type, then, as I have remarked, there could be no difference in the relative success of the two groups as measured by the two methods. Table 3.7 indicates the extent to which such cases do occur. It shows that 15 of the 44 subjects always treated all the cards of a particular type in the same way. Of the remainder 18 almost always responded in this way and 4 almost never did, the balance being made up of 3 subjects who belong, if anything, to

the former category and 4 subjects who belong, if anything, to the latter.

TABLE 3.7

NUMBER OF SUBJECTS IN EACH GROUP NOT TREATING ALL CARDS OF THE SAME KIND
IN THE SAME WAY ON DIFFERENT NUMBERS OF TRIALS (ALL STATEMENT-TYPES)

	<u>No. of Trials on which 'Incomplete' Response Made</u>															
	0	1	2	3	4	5	6	13	14	21	22	23	24	25
GR group	7	4	4	3	-	-	1		2	-		-	-	-	-	1
PR group	8	4	1	2	-	1	1		1	1		2	-	-	1	-
Total	15	8	5	5	-	1	2		3	1		2	-	-	1	1

It is not, of course, necessary to insist on a dichotomous classification of subjects who do not always treat all cards of a particular kind in the same way, but there are two different explanations for this type of response which suggest themselves and one of these appears to account rather well for cases in which an 'incomplete' response of this kind occurs on a relatively small number of occasions while the other appears more appropriate in cases where it occurs a relatively large number of times.

The first of these explanations is that subjects fail to notice or recognise all the cards of a particular kind. I have referred already to the fact that the 'five types of statement' task involves a perceptual component and this first explanation amounts to the suggestion that some 'incomplete' responses can be accounted for in terms of this component. It is a plausible explanation only in cases where this type of response occurs rather infrequently: it is simply incredible that subjects of the kind participating in the present research might repeatedly fail to notice one or more of the cards of a particular kind. In this latter kind of case it is much more plausible to assume that the subject does not realise that it is necessary to treat all cards of a given kind in the same way. The difference between the two kinds of subject can be stated most succinctly, perhaps, in terms of intention: the first kind of subject intends to turn over (or not turn over) all the cards of a particular type and accidentally fails to do so; the second kind of subject has no such intention. Finally, it has to be admitted that there may be 'mixed' cases where the subject starts off without the intention

but eventually comes to have it - or, of course, begins with the intention and later abandons it.

The importance of all this lies in its bearing on the question whether the third or the fourth method of assessing the respective success rates of the two groups is the more appropriate. If all 'incomplete' responses had been of the first kind, that is, if they had all been occasional rather than persistent, there might have been a case for simply ignoring them as representing 'noise' as opposed to 'signal' and using the third method of assessment. Since this is not the case, however, the fourth method has the advantage that it penalises the occasional incomplete response less than the third does, for it registers only one error in a trial where an error of this kind occurs as compared with, perhaps, two or three in a case where the subject intentionally treats all the cards of the kind in question in the wrong way. (On the third method, in contrast, both subjects would simply lose a point.)

On the other hand, the third method can be claimed to give a more accurate picture of the performance of the subject who persistently turns over only some of the cards of any particular kind. It may be said of such a subject that a fundamental condition of a correct solution has escaped him and that it is more appropriate that he should score nothing (as he might on the third method) than that he should receive (as he might on method four) as much credit as another subject who had grasped the fundamental point but given the wrong treatment to all the cards of a particular kind.

Since the 'occasional' cases are very much more common than the 'persistent' ones, and since both methods are equally fair to the subjects who never made an incomplete response, it seems that on the whole the fourth method is the more adequate. At the same time, some account ought to be taken of the results obtained by the third in view of the clarity with which it signals the failure of the eight persistent or semi-persistent offenders. It is perhaps also worth mentioning at this point that the scoring system based simply on answers to the question: 'Are (all) the cards of the four kinds given the appropriate treatment?' provides the basis for classifying, conveniently, if not entirely accurately, the responses made to the various statement-types, and this will be used in discussing the more general aspects of the 'five types of statement' task in the next section.

Our analysis of the differences between the two groups has become more and more detailed and it seems natural to take this process one stage further by tabulating the groups' scores on the two half-sets of each statement-type separately. This is done in Tables 3.8 and 3.9.

TABLE 3.8

AVERAGE SCORES OF THE TWO GROUPS ON ALL FIVE TRIALS, TAKING THE TWO HALF-SETS SEPARATELY FOR EACH STATEMENT-TYPE (MAXIMUM = 10)

	<u>Statement Type</u>									
	Ah		Ad		A		E		F	
	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.
GR group	9.0	4.3	7.3	7.4	9.2	4.5	9.0	8.1	5.4	7.0
PR group	8.9	3.2	5.9	6.1	8.5	3.1	8.5	7.5	3.4	6.6
Difference	0.1	1.1	1.4	1.3	0.7	1.4	0.5	0.6	2.0	0.4
Wilcoxon T		58.5	67.0	85.5		59.5			56.0	
N ranks		21	21	22		21			22	
p (2-tailed)		<.05	N.S.	N.S.		N.S.			<.02	

TABLE 3.9

AVERAGE ERRORS FOR THE TWO GROUPS ON ALL FIVE TRIALS, TAKING THE TWO HALF-SETS SEPARATELY FOR EACH STATEMENT-TYPE (MAXIMUM = 20)

	<u>Statement Type</u>									
	Ah		Ad		A		E		F	
	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.
GR group	1.2	10.7	4.6	4.5	1.0	10.9	1.5	3.2	8.3	5.6
PR group	1.5	13.0	7.0	7.1	2.0	12.9	2.3	4.4	12.5	6.2
Difference	0.3	2.3	2.4	2.6	1.0	2.0	0.8	1.2	4.2	0.6
Wilcoxon T		69.5	73.0	89.5		72.5			47.5	
N ranks		21	21	21		21			21	
p (2-tailed)		N.S.	N.S.	N.S.		N.S.			<.02	

Inspection of the scores and error totals of the two groups suggested that any significant differences that existed would be on the right hand half-sets of Ah, Ad and A or the left hand half-sets of Ad and

F. Accordingly, these were tested with the Wilcoxon. The difference on the left-hand half-set of F proved to be significant on both methods of assessment, the differences on the two half-sets of Ad and on the right hand half-set of A significant on neither, and the difference on the right hand half-set of Ah only on method three (Table 3.8). The Ah case was the least to be expected in view of my previous results. On the other hand, while the T value for this case is only just below that required for significance at the 5 per cent level, the corresponding value for the other form of the universal affirmative, A, is only .5 above it.

In view of the fact that these results for Ah and A occur only on the less reliable method of assessment little importance ought, perhaps, to be attached to them. On the other hand, if, as we came to expect on the basis of the 1928 paper by Wilkins, PR subjects are more prone than GR subjects to interpret the universal affirmative in such a way as to make it legitimate to infer its converse, and if the same is true of F statements, then the largest differences between the groups occur in exactly the half-sets we should expect. Hardly any subjects fail to recognise that a statement of the form, 'All cards with an X on one side have a Y on the other' requires the subject to turn over all cards with an X on their exposed sides. If he (of course mistakenly) supposes that this statement implies that all cards with a Y on one side have an X on the other, then he will also turn over cards with a Y on their exposed sides. The former response is made to cards in the left-hand half-set and is correct, the latter to cards in the right hand half-set and is wrong. Hence, if this mistaken assumption is commoner amongst PR than amongst GR subjects, the result should be a larger number of errors (of commission) for PR subjects on the right hand side and therefore, in terms of method three, a lower score on the right hand half-set. Finally, since the F type of statement is the mirror-image of a universal affirmative, the outcome in this case should be exactly the reverse, with PR subjects more frequently interpreting 'Only X's are Y's' as implying that all X's are Y's and so tending more often than GR subjects to turn over cards with an X (in the left hand half-set) which need not be turned.

In fact, as Table 3.10 shows, the PR group did have more errors of commission on the right hand half-set of Ah and A, and on the left hand half-set of F, but only in the last case did the T value even approach significance on a two-tailed test. It is, in fact, in errors of omission (also Table 3.10) that the difference between the two groups becomes sig-

nificant at the $p < .05$ level.

TABLE 3.10

MEAN ERRORS OF THE TWO KINDS ON THE RIGHT-HAND HALF-SETS OF Ah AND A, AND ON THE LEFT-HAND HALF-SET OF F, TYPES OF STATEMENT (MAXIMA RESPECTIVELY 8, 9, 12; 12, 11, 8)

	<u>Errors of Commission</u>			<u>Errors of Omission</u>		
	Ah (R.H.)	A (R.H.)	F (L.H.)	Ah (R.H.)	A (R.H.)	F (L.H.)
GR group	3.4	3.5	6.0	7.4	7.4	2.3
PR group	4.1	4.3	8.6	8.9	8.5	4.0
Difference	0.7	0.8	2.6	1.5	1.1	1.7
Wilcoxon T	90.5	95.5	62.5	85	73.5	58
N Ranks	22	21	21	21	21	21
p (two-tailed)	N.S.	N.S.	N.S.	N.S.	N.S.	<.05

It has to admitted that this was not an outcome that I had anticipated. What it means, of course, is that, faced with a statement of the form, 'Only cards with an X on one side have a Y on the other' PR subjects are significantly more likely than GR subjects to fail to recognise the need to turn over cards in the left-hand half-set which do not have an X on the exposed side. By itself this a response which it is difficult to interpret. What it perhaps suggests is that PR subjects are more likely, not merely to assume that 'Only S is P' implies that all S is P but that statements of these two forms mean the same thing. For this interpretation to be wholly convincing there would, of course, have to be a corresponding tendency for PR subjects to turn over cards of the named variety in the left-hand half-set, and we have seen that although this does happen, and does happen more often with PR subjects, there is no significant difference between the two groups in this respect.

It may be not unreasonable to suggest that the findings of the previous paragraphs may be interpreted as showing that PR subjects are more prone than GR subjects to confuse A and F types of statement, both, it seems, being read as an amalgam of A and F, equivalent, roughly, to 'All S is P and all P is S'. The implication of this view is that PR subjects are more likely than GR subjects to respond to A and F statements

in the same way. In terms of the experimental task this means that if, for example, a PR subject responds to the statement 'All cards with an X on one side have a Y on the other' by turning all cards with an X and all cards with a Y on their exposed sides, he will respond in the same way to the corresponding F statement.

The third method of assessment, as we have seen, is based on a consideration of the question whether each of the four types of cards has been given the appropriate treatment, and it is natural to record success and failure on each of the four card-types by writing a plus or a minus in the record-sheet. On the convention that the four types of card are considered in the order, cards with named characters in the left-hand half-set, cards in the left-hand half-set without named characters, cards with named characters in the right-hand half-set, cards in the right-hand half-set without named characters, it is possible to set out, as in Appendix F, a complete record of the responses of all subjects on each trial of each statement-type. Where the same treatment is called for for each of the four kinds of cards (as in Ah and A) the extent to which the subjects' responses to two statement-types are the same can be assessed by noting the number of occasions on which a plus in a given position in one of the statement-types is matched by a plus in the same position in the other.

In the case of A and F statements exactly the opposite holds: every response which is correct for a given category of card in one is wrong for that category of card in the other. (Cards with named characters in the left-hand half-set must be turned in A but must not be in F, for example.) This means, of course, that a plus in a given position in A must be 'matched' by a minus in F and a minus by a plus. If we use this as a basis for assessing the extent to which responses to A and F types of statement are more often alike in the PR group than in the GR group, there is one other complication to be borne in mind. This refers to subjects who make 'incomplete' responses (page 78 above). On the third method of assessment a plus is scored only if all the cards of a given type are treated in the appropriate way. A subject who responds by turning only some of the cards in the left-hand half-set with named characters in both A and F will, therefore, score a minus in both cases for in the former case he has failed to turn them all over and in the latter he has failed to avoid turning any over. A minus in both cases clearly represents identity of response in this special case and not, as

in other cases, a difference of response. In using the procedure described above for assessing the extent to which a subject's responses to A and F types of statement were the same I adjusted the 'agreement scores' of the eight subjects (Table 3.7) who persistently turned over only some of the cards of a kind accordingly - that is, by counting minuses representing incomplete responses to a given category of card in A and F as indicating similarity of response rather than difference. A less important adjustment of an opposite kind had to be made in the process of calculating the corresponding scores for Ah and A - for in this case it sometimes happened that a minus in one of the statement-types represented an incomplete response while the corresponding minus in the other statement represented a complete response of the wrong kind. Matching minuses in such a case clearly represented different responses and not the identity of response otherwise indicated in these statement-types. Table 3.11 presents the average A-F and Ah-A 'agreement scores' for the two groups, the differences between them being tested, as usual, by means of the Wilcoxon Signed-Ranks Test.

TABLE 3.11

AVERAGE A-F AND Ah-A AGREEMENT SCORES FOR THE TWO GROUPS (MAXIMUM=20)

	<u>Statement Types</u>		<u>Diff.</u>	<u>p</u>
	<u>Ah-A</u>	<u>A-F</u>		
GR group	17.4	9.5	7.9	<.001*
PR group	17.2	12.2	5.0	<.001*
Difference	+0.2	-2.7		
Wilcoxon T	120	73		
N Ranks	21	21		
p (two-tailed)	N.S.	N.S.		

*Sign Test

The Ah-A agreement scores, the difference between them, and the difference within each group between Ah-A and A-F agreement scores were calculated in addition to the A-F scores and inter-group difference, in which, of course, we are primarily interested, because it seemed likely that they would assist in the interpretation of any difference found between the A-F scores of the groups. To the extent that the PR group does treat the A and F types of statement as if they were equivalent to a

greater degree than the GR group does this might have been part of a quite general tendency to treat all five statement-types as if they were equivalent. Such a possibility could have been ruled out if we had found a significant difference between the groups on A-F and no difference, or of course a difference in the opposite direction, on Ah-A (which should be treated as if they were equivalent). In fact, however, such a clear-cut outcome was not forthcoming: the Ah-A difference between the groups is not significant, though it is in the direction which is incompatible with the view that PR subjects tend, quite generally, to react to all statement-types as if they were equivalent to a greater extent than GR subjects do; on the other hand, the intergroup A-F difference is also not significant. Finally, it has to be admitted that even if this difference had been significant we should have had to conclude that the extent to which PR subjects confuse A and F types of statement is relatively marginal in view of the highly significant intra-group differences in Ah-A as compared with A-F scores. The tendency in both groups to treat the Ah and A statement-types as if they were equivalent is so much more pronounced than their tendency to treat the A and F statements as if they were equivalent that it was possible to obtain a $p < .001$ even using the relatively low-powered Sign Test (Siegel, 1956).

This concludes my examination of the main body of evidence relating to the question whether PR subjects understand some of the types of statement which play an important role in the Valentine test, Section B less well than their GR counterparts. Subsidiary evidence of a rather problematic kind drawn from the special characteristics of the fifth trials of each statement type will be presented in section 3.5, after the presentation of some other, more general differences in the performances of the two groups and the discussion of some of the general points arising about the nature of the experimental task as so far described.

The evidence presented in this section seems clearly to suggest that there is no difference between the groups so far as the Ad (universal-disjunctive) and E (universal negative) types of statement are concerned. The strongest evidence of a difference between the groups relates to the A (universal affirmative in 'categorical' form) and F ('Only S is P') statement-types, with the difference in the former being significant on either the third or fourth method of assessment, and the difference in the latter appearing, with $p < .02$ on the fourth and, for the reasons given, probably superior, method. Rather surprisingly, there is virtually nothing to

suggest that the two groups differ in respect of their understanding of the other (hypothetical) form of the universal affirmative, the one significant difference between them being in the success - as measured by the less dependable third method - with which the groups tackled the right-hand half-set in this statement-type. This outcome is surprising, not only because A and Ah are both universal affirmatives (and so logically equivalent) but also because, as Table 3.11 makes clear, the responses all subjects made to them were highly similar (and suggest psychological equivalence).

There was some evidence of a rather tentative kind to suggest that the differences between the groups were located in a way which supports the hypothesis that the PR group was more prone than their GR counterparts to interpret 'All S is P' as implying 'All P is S' and 'Only S is P' as implying 'Only P is S' (i.e., 'All S is P'). However, differences in the incidence of errors of commission in the relevant half-sets which would have lent support to this conclusion, though in the right direction, were not significant. An (unpredicted) difference in errors of omission in the F case suggested that the PR subjects might be more prone than GR subjects simply to read an F statement as if it were the corresponding A statement. However, an attempt to compare the two groups with respect to the extent to which they responded to the two types of statement as if they were equivalent, though once again showing a difference in the predicted direction, failed to reach significance.

The overall picture presented by these results cannot be said to be unequivocal except in the sense that all differences between the groups, with the exception of the two noted in connexion with the clearly unsatisfactory first method of assessing the success of the two groups, were in favour of the GR group. At the same time, the significant differences, such as they were, do seem to combine to make some kind of sense and to suggest that there are differences between the groups which might be detectable using the present experimental task, perhaps presented in a slightly different way - as it was in a subsequent experiment with the same groups described in the next chapter.

3.3 Two more general points of comparison between the two groups

Reflection on the experimental task raised two more general questions about the performances of the two groups. It seemed natural, in the first place, to ask whether they differed in the extent to which they

changed their responses from trial to trial within each statement-type. One or two subjects were strikingly 'rigid' in their responses, not simply within statement-types but over the whole series of twenty-five trials. It would obviously have been of interest if there had been grounds for supposing that the two groups differed in a way which could have been interpreted as showing a tendency towards greater rigidity amongst PR subjects. A difference in the opposite direction, on the other hand, might have been interpreted in terms of a greater instability of response on the part of the PR group. Of the two the latter would have been the more interesting outcome for the present investigation for it might reasonably have been seen as an indication of uncertainty, on the part of PR subjects, about the meanings of the statements involved - whereas, on the other hand, rigidity of response might have been regarded as an aspect of the PR group's problem-solving behaviour in general.

In order to compare the two groups in this respect I calculated a 'changeability score' for each subject, this being equal to the number of different responses to each of the five types of statement taken together. (Types of response, for this purpose, were represented by patterns of pluses and minuses, as described on a previous page.) The result was almost exact equality between the groups, the average (with a maximum possible score of 25, of course) being 11.0 and 10.9 for the GR and PR groups respectively. We can therefore conclude with unusual confidence that there was no difference between the groups in the extent to which they changed their responses from trial to trial.

Consideration of the other rather general aspect of the performances of the two groups was prompted by the finding, noted earlier, that the universal negative (E) was the easiest type of statement for the two groups. The result is a puzzling one, as I have noted already, because Wason (1959, for example) and others have shown that negative statements are more difficult to cope with than the corresponding affirmatives. The question naturally arises, therefore, whether the present finding may not be an artefact of the experimental task - which in turn prompts a question about the aspects of the task which might make it uncharacteristically easy for subjects of both kinds to cope with the universal negative. In fact one possibility very readily suggests itself, namely, that in a state of uncertainty a subject is likely to make what Wason has recently (1969) called 'matching responses', i.e., turn over only the cards with named characters. This is, of course, precisely the kind of

response required for success in the card-turning task with the E type of statement, and if it were a common response for all statement-types, it might account for the surprisingly high success-rate with the universal negative as compared with the others. In fact, the response, as we shall see in a moment, is not a very common one and there are other reasons for doubting whether the explanation of the surprising outcome in the case of the E statement is to be found here. I shall discuss these other reasons in the next section when I come to consider the experimental task as a whole. For the present the important point is that the number of occasions on which PR subjects made matching responses was significantly greater than the number of occasions on which GR subjects did so.

To establish this I simply counted the number of plus-and-minus patterns of the appropriate kinds occurring in all statement-types except E. The reason for omitting E is that, as we have just seen, a matching response in this case is indistinguishable from a correct response, and it seemed important not to confuse cases in which a matching response might reflect understanding on the part of the subject with cases in which this possibility seems to be excluded. This distinction between the E case on the one hand and the other four statement-types on the other seems to be corroborated by the fact that, whereas, as we have seen, matching responses occurred significantly more often with PR subjects in the four cases where they are not appropriate, they occurred less often in the E type, where they are (the averages for PR and GR subjects being 2.4 and 3.3 respectively.) Once again some account had to be taken of the ambiguity of the minus in the case of subjects who made 'incomplete' responses, especially in the Ad type of statement, where a 'four-minus' pattern might represent either an incomplete or a matching response, the two being clearly different in the respect that, of course, a subject who makes an incomplete response does not turn over all cards with named characters.

The average matching-response scores on the four types of statement taken together (maximum possible = 20) for the GR and PR groups respectively were 2.6 and 5.7. With a Wilcoxon $T = 66$ and N Ranks = 22 the difference between the groups is significant on a two-tailed test with $p < .05$. The interpretation to be placed on this outcome is not clear. It may be, as suggested earlier, that a subject makes a matching response when he is uncertain about which response to make and that PR subjects were simply more often in a state of uncertainty. In that case, the difference between the groups would amount to no more than additional

evidence of the overall superiority of the GR group on this task. On the other hand, in the case of Ah, A and F types of statement what looks like a matching response is likely to commend itself to subjects who interpret these statements as stating a reciprocal relationship between the two classes of cards mentioned. It has been a main part of the argument of this chapter that this is something PR subjects may be more prone to do - at least in the A and F cases. Further evidence in support of the point will be presented in the next section but one of this chapter and in the chapter which follows.

3.4 The nature of the experimental task: a discussion with references to the work of Wason and others The focus of interest in the preceding sections has, of course, been on the extent to which the performances of the two groups differ, and this will continue to be the primary concern of succeeding sections and chapters of this thesis. In the present section, however, it seemed appropriate to review certain general features of the experimental task in the light of the performances of the two groups considered together rather than in opposition. At the same time, comparison will be made between my results and those obtained by Wason (1966, 1968, 1969) and Wason and Johnson-Laird (1969) and consideration given to the aspects of their subjects' performances which have seemed of special interest to Wason and to the theories developed by him to account for these aspects.

The purpose of this section is primarily to arrive at some understanding of the 'mediating processes' involved in successful and unsuccessful attempts at a solution to the experimental task with a view to determining the validity of the assumption, on which the research reported in this part of this thesis has proceeded, that a subject's degree of success in it reflects the extent to which he understands the various statement-types employed. One outcome of my own study which seems to have a bearing on this question has already been presented and commented upon. This concerns the relative difficulty of the various types of statement, the universal negative (E) proving for both groups to be easiest, the F type most difficult, with A, Ad and Ah being very much of a muchness in this respect for the PR group and Ad being second easiest for the GR group.

Between them, Wason and Johnson-Laird have used only the Ah, Ad and A types of statement. All that I feel able to say with confidence

about their subjects' comparative success on these three types of statement is that they found the disjunctive easier than either of the others. Of particular interest for a discussion of the nature of the experimental task is, I think, the finding, reported in the joint paper, that the disjunctive is easier, not only in the affirmative form in which I presented it, but also in the partially negative (and, to my mind, much more difficult) form in which it is logically equivalent to an Ah or A statement: 'Every card has a number which isn't Roman on one side or it has a letter which is capital on the other' (Wason and Johnson-Laird, p. 16). On page 19 the authors comment as follows: "The result of this experiment, when compared with those of Wason (1968), show that expressing implication as a disjunction, 'either not-p or q', makes it easier to grasp than expressing it as the conditional, 'if p then q'." To be specific, it is, in the view of the authors, easier to grasp the meaning of the above disjunctive than the corresponding conditional: 'If a card has a Roman number on one side it has a capital letter on the other.' I must say this seems to me to be highly improbable - especially in view of the fact that the evidence from Wason's later study and from my own would make it appear to be necessary to add that the disjunctive in question is also easier to understand than the corresponding universal affirmative in categorical form: 'Every card which has a Roman number on one side has a capital letter on the other.'

It is true that in using the above case to suggest that there may be something wrong with the assumption that the five types of statement task is in every respect a satisfactory measure of the extent to which the meaning of a statement is grasped I am setting a subjective impression against what appears to be an empirical test of the relative difficulty of the two (or three) types of statement. At the same time, Wason's own researches on the relative difficulty of affirmative and negative statements would lead one to expect the order of difficulty of the disjunctive and conditional forms of the above statement to be the opposite of what they are as measured by the present experimental task and I think it necessary to look for features of the task which may make it easier to arrive at a correct solution for some statement-types than for others, more or less regardless of the inherent difficulty of the statements themselves. In the previous section, for example, I considered the possibility that the surprising ease with which members of both groups could apparently cope with the universal negative might be explained in

terms of the fact that the correct choice of cards to turn over in the E case is identical with the cards turned over in a 'matching response', where the subject may be assumed to turn over the cards with the characters named in the statement more or less mechanically. As it happens, it does not seem likely to me that this particular accident of the experimental situation is likely to have played a major role in producing the rather unexpected relationships in the apparent difficulty of the five types of statement. For one thing, as I have already shown in the previous section, the average number of matching responses is quite small. Moreover, if subjects did tend to make matching responses on a preponderating number of occasions, we should expect the Ad type of statement to be the one on which most failures occurred - for the correct choice of cards in this case is the exact opposite of that involved in a matching response, while in the A and Ah cases the two selections at least overlap - whereas it is at least no more difficult than the A and Ah statements for the PR group and actually easier, if anything, in the case of the GR group.

Finally, in this connexion, it seems appropriate to refer to the very peculiar phenomena which have been the focus of Wason's research, and in particular the apparent inability of his subjects to apply to the problem of choice of card their recognition, in 'therapy sessions', that a card in the right-hand half-set of an A or Ah problem which does not have the second-named character on its exposed side may have the first-named character on its hidden side and so falsify the statement in question. It is, perhaps, significant that in his 1969 paper (p. 477) Wason has largely abandoned his earlier (and to my mind unconvincing) theories about the sources of such behaviour and seems to be relating it rather to characteristics of the task, in particular to the need to recognise cards without named characters as such, and the possible failure of subjects to understand the reciprocal nature of the relationship referred to in the expressions, 'on one side....on the other'. (Though the first of these points is surely still problematic, since it appears to apply to cards in the right-hand half-set of an Ah or A problem but not to cards in the left-hand half-set (which, of course, subjects overwhelmingly treat in the appropriate manner). More important, perhaps, it is, as far as I can see, inconsistent with the relative success subjects have with the Ad type of statement, for here all the cards which need to be turned are ones without the characters named in the statement.)

It seems very likely that the anomalous results so far as the relative difficulty of different statement-types is concerned are to be attributed to the operation of a multiplicity of factors. Perhaps, in conclusion, I should offer one other possibility which occurs to me in relation to my own results. This is the importance of symmetry as between the two half-sets. In Ad and E the correct treatment for the cards with named characters is the same in both half-sets - respectively, leave unturned and turn. In the other three statement-types, in contrast, the cards with named characters have to be turned in one half-set and left unturned in the other. The consequence, as the summary of correct responses on page 70 shows, is that it is relatively easy to formulate a rule for the Ad and E cases as compared with the others. Obviously, this feature by itself cannot explain the success with which subjects deal with these statement-types since the rule has to be the right one if the response is not to be totally wrong. It might be tempting to suggest that the matching-response hypothesis and the present one together account for the outcome at least in the case of GR subjects, for it is obviously easier to deal with cases where both rules are favourable (as in E) than with cases where only one is (as in Ad) while, of course, cases in which neither rule applies (Ah, A and F) are most difficult of all. But this theory, too, fails to fit all the facts, in particular the fact that the average number of matching responses is quite small in both groups.

Fortunately, for my purposes, the considerations presented above appear to cast doubt on only a part of the assumption, on which, as we have seen, the research reported in this part of this thesis depends, that the present task is a measure of the subject's understanding of the statements employed. This is that performance on the task reflects the relative difficulty of the various types of statement. Another part of the assumption, and the part most pertinent to the purposes of this thesis, is that within any one statement-type level of success is a measure of degree of understanding. Some of Wason's results may seem to cast doubt on this too: it might be argued, for example, following Wason's 1969 discussion, that the differences between the groups principally reflect differences in the extent to which their members function at the level of formal operations, or else the extent to which they 'regress' to the concrete level when faced with a task which is novel and abstract. The difficulty in maintaining the second part of the assumption about the relationship between success on the 'five types of statement' task and the extent of the subject's understanding of the statements used seems, nevertheless, to be

less than the difficulty in maintaining the first, and it will be apparent from the next section of this chapter and from the later experiment described in the chapter which follows that failure in the task is at least in part to be explained in terms of certain kinds of misunderstanding of the statements rather than wholly in terms of special features of the task. Indeed, the truth about the conditional and equivalent disjunctive (which appears to be more difficult to understand) may be partly, as I think Wason and Johnson-Laird suggest (pp. 19-20), that the complexity of the latter prevents subjects from making plausible but erroneous assumptions about the relationship between the relevant classes of cards which they are prone to make when these are stated in the simpler forms of an Ah or an A statement - and in particular, of course, the assumption that the relationship is a reciprocal one.

Finally, in this section, it may be of some interest to indicate the extent to which my results agree with those obtained by Wason and Johnson-Laird. To make this possible I have for the most part translated their results into my own terms, this being a relatively simple matter because their subjects virtually never made 'incomplete' responses - even when, as in the studies reported in the 1969 papers, there were two of each type of card. I thought it useful to present my own results in a different form from those hitherto used. This form does not indicate the relative difficulty of the various statement-types or the relative success of the two groups on the different statement-types. It has, however, the advantage that it makes it possible to detect any differences in the type of response favoured by the two groups (Table 3.12) and in the type and location of errors made by them (Table 3.13). The figures for the Wason studies always refer, of course, to the initial selections of his subjects and take no account of the selections they made after his various 'therapies'. For the Ah figures I have combined the results for Wason's (1968) experimental and control groups. The Ad figures are for the affirmative form of the disjunctive only. (The main difference between these and the partially negative equivalent of the conditional is that 7 out of 48 subjects seem to have been unable to cope with the negative in the antecedent: if these are added to the number scoring a complete success, the two patterns of choice are virtually identical, and either way, incidentally, the pattern of choice for this equivalent of the conditional is quite different from the quite distinctive pattern for Ah and A.)

Down the left-hand side of Table 3.12 are listed all possible patterns of response in terms of success (plus) and failure (minus) on the four components of a correct response (corresponding to the four types of card in the order given in the previous section, p. 83, and therefore to Wason's P, \bar{P} , Q, and \bar{Q} .) Opposite them is the percentage (to the nearest whole number) of all responses of the group in question to that statement-type represented by responses of the pattern in question. Three additional points ought to be made. These are, first, that the percentages given for the Wason groups refer to single responses of different subjects whereas the percentages for the GR and PR groups refer to the responses of my 44 subjects on the five successive trials.¹ The N's on which the percentages are based are, therefore, 110 for all GR and PR groups and 60, 48 and 32 for Wason's Ah, Ad and A groups respectively. Secondly, I did not think it necessary, on this occasion, to distinguish cases where a minus represents an incomplete response from those in which it represents a complete, but incorrect, response. The incomplete responses (in which, it will be remembered, all the cards of a given type are not treated in the same way) will, of course, tend to appear in the four-minus row at the bottom of the Table or to be scattered throughout the predominantly minus rows immediately above. Finally, it is perhaps worth reiterating the point that a given pattern of pluses and minuses will indicate a different set of component responses in different statement-types (except in Ah and A) and, conversely, a given set of component responses will be represented in different statement-types by different patterns of pluses and minuses - so that a matching response, for example, is represented by ++-- in Ah and A, ---- in Ad, ++++ in E and --++ in F.

Inspection of the Table prompts the following tentative remarks.

- (1) On Ah and A Wason's subjects appear to have scored rather fewer complete success/^{es} than even my PR group. This is, however, one respect in which the 'one off' character of the Wason results is misleading, for only one subject out of my 44 achieved complete success on Ah and A in the

¹ It might be argued that for strict comparability I should have used only the first trial responses of my subjects. On the other hand, my subjects were given no information between trials which could have encouraged them to change their responses, pooling their responses presumably makes the overall distribution more representative of the groups' views, and, inasmuch as I am interested in a comparison between these groups as well as between them and Wason's, the pooled results are therefore to be preferred.

TABLE 3.12

PERCENTAGES OF RESPONSES OF GR, PR AND WASON GROUPS WHICH WERE
OF THE VARIOUS POSSIBLE PATTERNS

Pattern	<u>S t a t e m e n t</u> <u>T y p e s</u>												
	Ah			Ad*			A			E		F	
	GR	PR	Wason	GR	PR	Wason	GR	PR	Wason	GR	PR	GR	PR
++++	13	11	5	60	40	75	14	9	-	66	48	20	9
+++-	37	31	26	-	1	-	42	26	50	5	2	2	1
++-+	18	4	6	2	-	-	14	4	6	9	20	6	4
+---	1	-	-	-	4	-	-	-	-	1	1	6	4
----	-	-	-	-	2	-	2	1	-	5	7	19	6
++--	17	35	50	-	-	4	21	37	41	5	5	4	3
+--+	2	6	5	-	-	-	1	3	-	1	-	-	1
-+-+	-	-	5	1	8	-	-	1	3	1	3	-	-
---+	1	1	-	18	19	-	-	-	-	3	6	5	8
----	-	-	-	2	-	-	-	1	-	-	3	11	26
-++-	1	1	-	1	1	-	1	4	-	-	-	12	15
+---	1	3	2	1	1	-	-	2	-	-	2	1	1
-+--	4	6	-	1	1	-	1	8	-	1	1	2	2
---+	-	-	-	2	2	-	1	2	-	2	2	1	6
----	-	-	-	1	1	-	1	2	-	1	-	5	11
----	6	4	-	12	21	19	4	1	-	3	1	8	6

* The response of one of Wason's Ad subjects is not included because its nature is not clearly enough described to allocate it to any of the above patterns.

first trial (compared with 3 out of Wason's 60 Ah subjects). Otherwise, there is clearly a large measure of agreement amongst all groups on the most popular patterns of response, viz., +++- (what I shall call the 'half-matching' response, turning only the cards with the characters first named - Wason's P cards) and ++--, the matching response. There is some evidence to suggest that the matching response occurs rather less often with the GR group and that the +-+ response occurs more often. This latter point is of some interest in that it ^{may reflect} what may be an excessive caution on the part of this group, for in this pattern of response, as in the corresponding one in F, the difficult component of a correct response is dealt with appropriately and only an error 'on the safe side' remains.

(2) On Ad Wason's group achieved complete success strikingly more often than my PR group and noticeably more often than the GR group. Offsetting this, partly, is the incidence, in GR and PR groups, of responses (-+-) which consist in turning over every card. I remarked in a previous section that this is the correct response if the disjunctive is interpreted in the exclusive sense, and although none of my subjects adopted this interpretation consistently, the incidence of -+- responses revealed in the

above Table may reflect a, perhaps to be expected, oscillation between the two possible interpretations. Presumably, the absence of any responses of this kind in Wason's group is to be attributed simply to chance, for in their paper he and Johnson-Laird report (p. 18) that 'three subjects gave evidence of reasoning exclusively'. Apart, perhaps, from the 8 per cent of PR responses which consisted in turning none of the cards over (+--+), the only other percentages of any size in Ad are those for the ----, matching response (though this seems to have been somewhat less popular with the GR group than with the others).

(3) On E the relatively high incidence of completely correct responses is accompanied by a very wide distribution of the remaining responses over the other possibilities, with only the half-matching response (++-+) accounting for a substantial proportion of the total.

(4) Responses are also widely dispersed in F and an additional feature of the distribution here is that differences in the preferences of the two groups are more apparent than they are anywhere else. The one incorrect pattern in which there is a noticeable preponderance of GR, as opposed to PR, responses is -+++, the pattern referred to in the previous paragraph in connexion with the remark that the GR group may show a tendency to excessive caution. There are substantially more matching responses (-+++) from the PR group and half-matching responses also make up a somewhat larger proportion of their total. The other response which accounts for ^a/sizeable proportion of the responses of both groups is -+--, which consists in turning over all the cards in the left-hand half-set and the cards (in the right-hand half-set) with the characters named second.

(5) Finally, it may be as well to admit that the fact, noted under the two previous points, that the distribution of responses is greatest in E (the easiest statement, whether we consider errors or time taken) and in F (the most difficult, again on either criterion) is something for which I can think of no adequate explanation. By itself, of course, the F case presents little problem: as subjects find it the most difficult we might expect them to try a wide variety of incorrect responses. The same explanation obviously will not do for E.

In Tables 3.13 and 3.14 the performances of the two groups are analysed in terms of the location of the errors they made and the extent to which the errors they made in the two half-sets were errors of omission or (by inference) errors of commission. It is possible to present comparable results for Wason's groups because, of course, the division of

my cards into two half-sets lying to right and to left of each other is a relatively unimportant feature of my experimental set-up: as I have pointed out, a more or less straightforward translation of Wason's 'P, \bar{P} , Q, \bar{Q} ' terminology into my terms is always possible. Similarly, an estimate of the number of errors Wason's subjects made 'in each half-set' and also of the proportion of these which were 'errors of omission' can readily be made on the basis of the information he gives about the cards which his subjects turned over. Indeed, a rough approximation to the truth can be read off Table 3.12.

The results for my own groups are based on the error scores found in the course of applying the fourth method of assessment. As the intention in this section is to make comparisons between statement-types as well as between groups, it was necessary to adjust the raw error scores to take account of the fact that there were not always two cards of each kind in each half-set and therefore not always equal arithmetical probabilities of the two kinds of error. The procedure I adopted to effect this adjustment was to divide the actual number of errors of omission made by each group in each half-set on each trial in each statement-type by the number possible and multiply the result by two. Total error scores were then adjusted by subtracting the difference between the raw EO score and the adjusted EO score in each half-set/trial from the raw total error score. As a result, the arithmetical probabilities of the two types of error may be assumed to be the same for all half-sets and all statement-types and any differences in the proportions of them actually found may be taken to reflect the special problems presented by the various statement-types.

TABLE 3.13

ADJUSTED FREQUENCY AND PERCENTAGE OF ERRORS OCCURRING IN THE LEFT-HAND
HALF-SET

	<u>S t a t e m e n t</u>			<u>T y p e</u>	
	Ah	Ad	A	E	F
GR group: %	10	52	8	30	63
N	(24)	(99)	(19)	(29)	(192)
PR group: %	11	50	13	32	71
N	(32)	(145)	(39)	(44)	(294)
Wason: %	8	45	2		
N	(7)	(36)	(2)		

There is clearly a very large measure of agreement between all three groups in the way in which errors are distributed between the half-sets of the various statement-types. Errors in Ah and A are overwhelmingly in the right-hand half-set, a fact which may be taken to support or to illustrate Wason's point that the problem about these statement-types centres on the Q and \bar{Q} cards. The even distribution of errors between the two half-sets of Ad is what one might expect given its 'symmetrical' character, already referred to. On the other hand, one of the surprising features of Table 3.13 is that no such even distribution of errors occurs in the other symmetrical statement-type, E: the two groups agree closely in placing rather more than two-thirds of all errors in the right-hand half-set. It seems possible to offer an explanation for this outcome, and for the fact that the distribution of errors in F is not as extremely one-sided as in Ah and A (of which it is in other respects the mirror-image) in the light of the results presented in Table 3.14 where the division of errors into their two types is given for each half-set.

TABLE 3.14

ADJUSTED FREQUENCY AND PERCENTAGE OF ERRORS IN EACH HALF-SET
WHICH WERE ERRORS OF OMISSION

	<u>S t a t e m e n t</u> <u>T y p e</u>									
	Ah		Ad		A		E		F	
	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.	L.H.	R.H.
GR group: %	54	64	32	35	58	65	69	57	30	58
N	(13)	(134)	(32)	(33)	(11)	(148)	(20)	(39)	(57)	(64)
PR group: %	44	65	47	50	56	64	61	67	35	51
N	(14)	(168)	(68)	(71)	(22)	(171)	(27)	(62)	(107)	(63)
Wason: %	--	57	50	50	--	67				
N*	(0)	(50)	(18)	(22)	(0)	(60)				

* In this table, and in the previous one, the error scores for Wason's groups in Ad and A have been doubled, in comparison to his Ah group, to take account of the fact that there were two cards of each type in the former cases and only one in the latter.

The degree of agreement between the groups shown in Table 3.14, though not so large as in the previous table, is clearly still very considerable. In percentage terms, the largest discrepancy is in the left-hand half-sets of Ah and A where about half the errors in my groups were EO's while none of the other group's were. In terms of frequencies, however, the difference is very much less and may perhaps be attributed

to a combination of 'incomplete' responses and to perceptual errors. The other main discrepancy is in Ad where the GR group shows a clear preponderance of EC's in both half-sets whereas the division of errors into the two kinds is nearly even in the other two groups. In the absence of any better explanation, this result may be taken as confirming a tendency, noted earlier, for the GR group to err on the side of turning over too many, rather than too few, cards.

Comparison between the two groups apart, Table 3.14 is perhaps best read in conjunction with Table 3.13. It is then possible to see that roughly two-thirds of errors in the right-hand half-sets of Ah and A (where about 90% of errors occurred) were errors of omission, failure to turn over what Wason calls the \bar{Q} cards, the cards in that half-set without the character named. Table 3.14 shows that the majority of errors in both half-sets of E were EO's while more than two-thirds of the errors in the left-hand half-set of F were EC's. These facts may hold the key to the two problems left over from Table 3.13, the uneven distribution of errors over the two half-sets of E, and the absence of any extremely one-sided distribution of errors in the two half-sets of F of the kind found with Ah and A.

My suggestion is that subjects tend to regard all statement-types except Ad as being simply about the class of cards with the first named characters. Ad is an exception because the grammatical structure of the sentence in this case fails to 'attach' either of the named characters to the subject-term. In A, E and F the grammatical subject is of the form, 'cards with an X on one side' and in Ah, where the grammatical subject is simply 'a card', this occurs in the same subordinate clause as the expression for the first named character. In Ad, in contrast, 'every card' is said to have 'either X... or Y...'. .

Such a tendency to suppose that the statements other than Ad refer simply to the class of cards with the first named character (Wason's P cards) would, if it were real, lead subjects to neglect cards with the character named second (Q cards). Now this is obviously not an overriding tendency for we have already seen that subjects do tend to turn over Q cards in Ah and A and it runs counter to the tendency to make matching responses. On the other hand, it does help to explain why 'half-matching' responses, when they occur, tend very much to consist of 'matching' the first half and hardly ever of matching the second. (Since the patterns

for 'second-half matching' are --- for Ah and A, +--- for AD, -+++ for E and +-+- for F, it is possible to see from Table 3.11 that the percentages of such responses were 4, 1, 1, 5 and 0 respectively for the GR group and 6, 1, 8, 7 and 0 for the PR group, as compared with 'first-half matching' percentages of 37, 2, 42, 9 and 5 (GR) and 31, 2, 26, 20 and 11.)

In the absence of any bias emanating from this source in the case of Ad there would be no tendency to make errors of omission (or, therefore, errors of any kind, so far as one can judge) in one half-set rather than another. In E, on the other hand, to the extent that the above effect is operative EO's could be expected to fall predominantly in the right-hand half-set, as we have seen they do, and so upset the even distribution of errors as between the two halves which I have suggested we should expect in the 'symmetrical' statement-types. The same factor would help to keep down errors of omission in the left-hand half-sets of Ah and A but tend to raise the number of EC's in the same half-set of F, in this way reducing the extreme one-sidedness of the distribution of errors as between the two half-sets which we might otherwise have expected in view of the relationship between F and Ah and A. In fact, as Table 3.14 shows, about two-thirds of the errors of both groups on the left-hand half-set of F are errors of commission.¹

It may be apparent at this juncture that at least some of the responses made by subjects in the 'five types of statement' task are 'over-determined' - probably in fact, but certainly in terms of the hypotheses developed in this chapter. In the case of F statements, for example, the tendency for subjects to make errors of commission in the left-hand half-set has been attributed to at least three different factors: the tendency to turn over cards with named characters, more or less without regard to the statement-type, the tendency to suppose that F statements are 'about' only the class of cards with the first named characters, and the tendency to suppose that 'Only S is P' implies that all S is P.

¹ One is naturally prompted to look for some way of manipulating this effect experimentally. One which suggests itself, in the case of Ah, is changing the order of the clauses so that the statement reads, 'A card has an X on one side if it has a Y on the other'. One would expect, in this case, to find an increase in EC's in the (former) left-hand half-set and of EO's in the (former) right-hand half-set (where the P cards would no longer benefit from the effect). Perhaps this is why Hughes (1966) is reported by Wason (1968, p. 281) to have found that "the logically equivalent expression, 'Q if P', causes, if anything, even more difficulty" than the normal, 'if P then Q'.

Now in fact, as I have suggested in passing, all these factors may play a part in producing the patterns of errors observed, but the important question, for my purposes, is about the extent to which the last is real and important, for it alone seems able to provide us with a clue to the origins of error in syllogistic reasoning. To answer this question, however, we appear to need a slightly different experimental set-up - for example, of the kind described in the next chapter and anticipated, to some extent, in the special features of the fifth series of trials, in the present experiment, described in the next section.

3.5 The fifth series of trials My consideration of this series will be brief because, as will shortly be apparent, there were serious difficulties in drawing any conclusions about certain aspects of the subjects' responses, and these were eliminated in the experiment to be described in the next chapter.

The special features of this fifth trial have already been described, at the end of the previous chapter. Very briefly, what happened was that, after the subject had indicated, on the strip of paper bearing the diagrammatic cards, which cards he thought needed to be turned over, the experimenter produced a set of actual cards with the appropriate characters on their exposed sides. Subjects were then asked to turn over the cards they had marked and to say whether the statement was in fact true or false of that set of cards. If they thought it false, then they were to indicate, by drawing a ring round the appropriate tick on the paper strip, which card or cards made it false. Figure 3.1 shows the characters on the cards (characters on the hidden side in brackets) as well as the statements which accompanied the cards

FIGURE 3.1

STATEMENTS AND CHARACTERS ON THE CARDS IN THE FIFTH SERIES OF TRIALS

Ah statement: If a card has a heart on one side, it always has a cross on the other.

Cards: ♡(+), ♡(+), ◇(□), ♠(△), +(♡), □(♡), △(♠), □(♠)

Ad statement: Every card has a capital letter on one side or else a club on the other.

Cards: e(♣), A(◇), t(♠), m(♠), ◇(E), ♡(T), ♠(A), ◇(M)

FIGURE 3.1 (CONTD.)

A statement: All the cards with a heart on one side have an even number on the other.

Cards: ♡(2), ♠(4), ♣(7), ♡(8), 3(♥), 5(♠), 7(♣), 8(♥)

E statement: No cards which have a spade on one side have a Greek letter on the other.

Cards: ♠(β), ♠(t), ♣(a), ♣(e), ♠(α), ♠(ϕ), ♡(τ), ♣(σ)

F statement: Only cards with a vowel on one side have a diamond on the other.

Cards: a(♥), e(♠), o(♣), z(♥), ♠(a), ♠(e), ♡(o), ♣(z)

In subsequent references to the cards in these sets I shall regard them as numbered from 1 to 8 from left to right. Thus it will be apparent that card 6 makes the Ah statement false and card 5 the A statement false. The other three statements are, of course, true.

The original, and rather simple, aim in including this task in the present experiment was to answer the natural question about how many subjects would actually come to the right conclusions about the truth-values of the five statement-types, given their choice of cards. It seemed that a number of interesting additional possibilities would also be touched upon at least - amongst these being the effect of having to deal with real cards as opposed to diagrams of cards, the success with which subjects could draw the appropriate inference upon discovering what was on the other side of a card, and so on. In his 1966 contribution Wason had said that subjects 'hardly ever' thought that a card with a consonant on one side and an even number on the other falsified the Ah statement: 'If a card has an even number on one side, it has a vowel on the other'. I was particularly interested in the possibility of this kind of error because of its relationship to the confusion, to which I have frequently referred, between the A and F type of statement. It will be seen from Figure 3.1 that a person who thought that 'All S is P' implies that all P is S might be expected to say that the A statement is false, not only because of card 5 but also because of card 2, while a subject who supposed that 'Only S is P' implies that all S is P might think that the F statement is false because of cards 1, 3 and 7. In fact no one turned over card 2 in A and I had to wait for the later experiment to get an answer to

the question whether this particular misunderstanding showed itself in this way, and whether there was a difference between the two groups in this respect. So far as the first of these questions is concerned, however, the same did not appear to be true of F - as we shall see in a moment.

So far as the answer to the original, simple question is concerned, the average number of correct verdicts about the truth-values of the five types of statements in the fifth trial proved to be 3.3 and 2.6 for the GR and PR groups respectively, a difference in the expected and, by now, familiar direction - and, of course, to some extent the product of the GR group's superiority in the matter of choosing cards to turn over. The difference is not a significant one and is anyway one on which it would be rather unwise to place any weight since eight of the correct verdicts reached by either group were based on inadequate or misunderstood evidence: in fifteen cases subjects concluded (rightly) that a statement was true although they had not turned over all the cards which might have proved it false, and in one case the subject declared a statement to be false on the basis of a card which did not in fact make it false while ignoring another card which really did make it false.

Of more interest, clearly, are those cases in which subjects drew the wrong conclusion in the face of all the relevant evidence. There were 13 and 22 such cases in the GR and PR groups respectively, distributed amongst the statement-types as shown in Table 3.15.

TABLE 3.15

FREQUENCY OF WRONG VERDICTS ON THE TRUTH-VALUES
OF THE FIVE TYPES OF STATEMENT

	Statement Type				
	Ah	Ad	A	E	F
GR group	--	6	1	2	4
PR group	1	4	1	5	11
Total	1	10	2	7	15

Since the Ah and A statements were actually false, mistakes about these consisted in failing to recognise that the critical card in each case was an exception to the rule. In view of the small number of

such cases it may seem enough to attribute the failure to inattention on the subject's part or to other chance factors. On the other hand, the critical cards are in each case in the right-hand half-set, with the result that the character first named is on the reverse side, and this prompted the question whether something similar to the 'directional' effect discussed by Wason (1969) might not have been operating. The two effects would not be quite the same, of course, for the one referred to by Wason is at the level of thought and is intended to explain why subjects fail to realise that there might be a P on the back of a \bar{Q} card (to use a shorthand based on Wason's designations), whereas I am looking for an explanation for something much more like a perceptual failure on the part of the small number of subjects who failed to recognise a $P\bar{Q}$ case when the P was on the reverse side. As we shall see, there is some further evidence of a casual kind in support of this possibility in connexion with mistakes made with the Ad statement. Experimental evidence could obviously be obtained by asking subjects to detect exceptions to a rule when these exceptions are presented in the normal and 'reverse' position, and although it was not regarded as an important part of the purpose of the experiment to be described in the next chapter, some additional evidence of a rather more substantial kind on this point will be presented there.

Since the Ad, E and F statements of the fifth series were true, mistakes here consisted of wrongly supposing that one or more cards were exceptions to the rule stated. Table 3.16 shows, for each statement-type, which cards were viewed in this way and by how many subjects.

TABLE 3.16

NUMBER OF SUBJECTS WRONGLY REGARDING CERTAIN CARDS AS FALSIFYING Ad, E AND F STATEMENTS

	S t a t e m e n t T y p e											
	A				E				F			
Card Nos.	2	5	6	8	3	5	7	8	1	2	3	7
GR group	4	1	-	2	-	-	-	1	2	1	-	1
PR group	2	1	1	1	2	1	1	2	7	1	7	4
Total	6	2	1	3	2	1	1	3	9	2	7	5

The F case is the clearest. Cards 1, 3 and 7, as we have seen,

are all of a kind which would falsify the A counterpart of the F statement in question, and it seems reasonable to conclude that subjects who thought these cards were exceptions to the rule contained in the F statement suppose that 'Only S is P' implies that all S is P. The difference in the frequency with which this misunderstanding occurred in the two groups is in the expected direction but is not, of course, significant as it stands, and is anyway to be attributed partly to the fact that fewer GR subjects turned over these cards and fewer were therefore faced with the question whether they were exceptions to the rule. The experimental set-up employed in the later experiment eliminated this source of uncertainty too - although a conclusion could be drawn only in the A case, as we shall see.

Most of the mistakes in the Ad case can also be attributed to a single confusion, though this time one which was unexpected. Cards 2, 5, 6 and 8 all bear one, but only one, of the named characters. It seems, therefore, that subjects who said that these cards proved the statement false must have been either misreading it or else misconstruing it as a universal affirmative. I find it hard to take the second alternative seriously since it involves one in the assumption that these subjects can not distinguish a universal-disjunctive from a universal affirmative, and this is a mistake which may occur at an early stage of development (Watts, 1944) but is surely rather unlikely to occur in adults of very superior ability. The other possibility which suggests itself - that subjects misread the statement - requires us to assume that they mistook the statement, 'Every card has a capital letter on one side or a club on the other' for the statement, 'Every card which has a capital letter on one side has a club on the other' (the form of the universal affirmative favoured by Wason in his 1969 paper). Of the two alternatives this seems to me to be the more plausible, especially on the reasonable assumption that subjects were by this stage beginning to be tired and perhaps less attentive than at earlier stages of the experiment.

A more alarming possibility, of course, is that Ad statements were consistently misread (or misconstrued) by these subjects as universal affirmatives. I tried to check this by noting the extent to which the responses of the eight subjects in question were the same for Ad and A types of statement over the five trials. The agreement scores¹ (max.=20)

1 For the method of calculating these scores see above p. 83.

for Ad and A ranged from 3 to 15 with an average of 9.9, as compared with a range of 9 to 20 and an average of 16.4 for the Ah and A agreement scores for the same small group. In seven of the eight cases the extent to which these subjects treated Ad and A statements as if they were equivalent (as measured by the agreement scores) was less than the extent to which they treated Ah and A statements as if they were equivalent - as, of course, these are - the remaining subject having a score of 9 in both cases. The difference between the scores on the two pairs of statements is, then, highly significant ($p = .008$, Sign Test, two-tailed) and although of course it is not possible to say with any confidence that the misreading (or misconstruction) of the Ad statement as an A statement occurred only in the fifth trial, it does seem legitimate to conclude that the confusion was not a frequent one.

One final comment about these mistakes on Ad concerns the fact, revealed in Table 3.16, that card 2 (in the left-hand half-set) was recognised as an exception to Ad, misread as A, as often as the other three cards (in the right-hand half-set) taken together. As card 2 was turned only 7 times by the subjects in this group, while the others were turned altogether twice as often (4, 6 and 4 times, respectively) this may perhaps be taken as additional, casual (and, of course, inconclusive) grounds for considering further the possibility that some kind of directional effect makes it relatively more difficult for subjects to recognise cards as exceptions to a rule if the first named character is on the reverse side.

It is much more difficult to find a single explanation for the majority of mistakes on E than it has seemed to be in F and Ad. Inspection of the cards erroneously supposed to falsify the E statement suggests no more general possibility than that some subjects confused spades with clubs: this would account for the three cases in which card 8 was chosen. No explanation other than inattention suggests itself for the remaining four cases.

Lastly, in this section, a brief word should, perhaps, be said about the apparent effect on subjects of having to deal with actual cards as opposed to diagrams of cards. The most important observable way in which any effect might be expected to show itself is in changes in subjects' views about which cards they ought to have turned. If the effect had been a major one, for example, subjects might have shown a dramatic

change from failure to success, realising, when confronted with the problem in 'real' terms, exactly which cards they ought to have turned over. In fact, of course, Wason's 1966 paper made such an outcome extremely unlikely. On the other hand, some subjects did change their minds about the cards needing to be turned over at some point after the actual cards were produced, and I allowed them to turn over additional cards and record the fact of having done so by writing a tick with a plus in front of it under the appropriate card diagram. These were not, of course, taken into account in the comparisons between groups made in previous sections of this chapter but they do provide a partial indication of the extent to which the problem changed its character when subjects were faced with real cards.

The facts are as follows. Five subjects from each group turned over at least one additional card in one statement-type; three GR subjects and two PR subjects turned over at least one additional card in two statement-types; the remaining twenty-seven subjects made no change in their original choices. The number of subjects from both groups turning over additional cards in the five statement-types was 3, 2, 6, 3 and 6 for Ah, Ad, A, E and F respectively. Of the 16 and 23 additional cards turned by GR and PR subjects respectively 11 and 15 were in fact cards which ought to have been turned over. If we can exclude the possibility that these changes are due simply to a 'second thought', the above figures suggest some small degree of improvement in performance when actual cards are used. In the experiment to be described in the chapter which now follows subjects worked with real cards in all trials.

3.6 Review of the main points The points of central interest that have been established in the foregoing sections have, of course, related to the differences between the two groups. Stripped of any element of interpretation these were: (1) that PR subjects were significantly poorer at selecting the cards which needed to be turned over in A and F statements; (2) that PR subjects made significantly more mistakes in the right-hand half-set of A and the left-hand half-set of F; (3) that PR subjects made significantly more errors of omission in the left-hand half-set of F; (4) that PR subjects made significantly more 'matching responses' in the four types of statement other than E.

Even at this level these results are not without their problematic aspects. In particular it must be regarded as something of a myst-

ery that the significant differences noted above with respect to the A type of statement have only one, rather dubious, counterpart in the case of Ah, despite the fact that the two statement-types are not only logically, but also, in terms of the similarity of the response made to them by both groups of subjects, psychologically equivalent. But the assumption has been that the differences under (1) can be interpreted as meaning that PR subjects understand A and F statements (but not Ah, Ad or E statements) less well than their GR counterparts. We have seen reason to doubt whether success on the card-turning task can be taken to indicate understanding of statements when comparisons are made within groups and between statements (see also a later paragraph of the present section); fortunately, however, the assumption crucial to this research, that differences between groups within statement-types may be taken to indicate differences in the extent to which statements are understood by the members of the groups, appears to stand.

Granted, then, that the differences noted under (1) above mean that PR subjects on the whole understand A and F statements less well than GR subjects, it is natural to go on to ask what kind of misunderstanding is involved. Partly following Wilkins (1928) the hypothesis was that it would consist in assuming that A statements imply their converses and that F statements do the same - that is to say, that they state a reciprocal and not simply a one-way relationship between ^{the} classes represented by their terms. The differences under (2) above might be interpreted as supporting this hypothesis. Unfortunately, the two groups did not differ significantly in the extent to which they made the kind of errors in the two half-sets which one would expect on the basis of the hypothesis in question. On the contrary, the difference noted under (3) suggests a rather more radical conclusion in the case of F, namely, that PR subjects tend, to a greater extent than GR subjects, to read 'Only S is P' as 'All S is P'. More disturbing, scrutiny of the general pattern of responses suggested that the difference under (2) might be explained in terms of a greater tendency on the part of the PR group to make matching and half-matching responses. (See the difference under (4) above.) It seemed clear that we need some other way of deciding whether the difference in the adequacy with which the two groups respond to A and F types of statement is to be interpreted in the way described at the beginning of this paragraph. In section 3.5 we have seen one way in which this might be achieved, and in the chapter which follows this will be used in the case of A statements to establish the point beyond further doubt.

Some of the analyses of the present chapter have made it possible to draw conclusions about the nature of the experimental task which have a significance and interest quite apart from their bearing on the particular concern of this part of this thesis. It would perhaps be appropriate to conclude this chapter with a summary of the two most important of these.

The first has already been touched upon in passing. It is that success in the card-turning task with different statement-types does not appear to reflect in any straightforward fashion the relative difficulty subjects have in understanding the types of statement in question. The relative ease with which my subjects coped with the universal negative and the fact, discovered by Wason and Johnson-Laird (1969), that a statement of the form, 'Everything is either not-X or else Y' produces fewer mistakes, on the present task, than the (logically) equivalent hypothetical, 'If a thing is X, then it is also Y', represent the chief grounds for this conclusion. We seem forced to conclude that success as between statement-types is to a considerable extent due to the special features of the task. It is difficult to be certain what these special features are, but four which suggest themselves are as follows: (1) the extent to which success depends on turning cards with named characters; (2) the extent to which the treatment given to the two half-sets in a correct selection is the same; (3) the extent to which the form of statement encourages the view that it is a statement only about the class of cards with the character first named; and (4) the extent to which the form of the statement alerts subjects to the difficulties of the task. The first two of these factors may explain why my subjects found the E type of statement easiest to deal with, the last two why Wason and Johnson-Laird's subjects made fewer mistakes on the negative-disjunctive than on the affirmative-hypothetical form of the universal. In any case it does seem of some importance to try to devise experimental arrangements in which the operation of these different factors can be studied.

The other point of general importance is that, within certain very broad limits, the distribution of responses (and hence the location of errors and the incidence of errors of the two types) amongst undergraduates appears to be highly predictable. Wason has developed a number of hypotheses to account for the persistence of certain erroneous types of response, especially the very common errors of omission in the right-hand half-set of Ah and A. I have suggested that his latest hypothesis (1969)

according to which this type of error is due to a failure to recognise the 'reversible' nature of the relationship implied by the expression 'on the other side of the card', may have a counterpart at the perceptual level, my best evidence for this being due to be presented in the chapter which follows. So far as the general similarity of response between the undergraduates is concerned, I should expect further light to be shed on this if the task were presented to much younger subjects - say from the age of seven years upwards. It would be of particular interest to see if the distribution of the most popular responses varied very much over so wide a developmental span

CHAPTER FOUR

THE 'FOUR TYPES OF STATEMENT' EXPERIMENT

Summary All but one of the subjects in the experiment described in the previous chapter returned twelve to eighteen months later to take part in a modified version of the card-turning task. The purpose to be served by this subsequent experimental session was to provide answers to the following questions. (1) How stable are the responses of subjects in the card-turning task over a relatively long interval? (2) In the special circumstances of the later experiment what evidence is there of learning over a number of trials? (3) Are the differences between the groups which were significant in the earlier experiment still significant a year or more later? (4) If they are, is there also evidence of a fairly conclusive kind that some of these differences are to be attributed to a tendency for PR subjects to interpret the universal affirmative as if it implied the truth of its converse more often than GR subjects? (5) Is there any further evidence to support the view that exceptions to a rule are less easily recognised as such when they occur in the right hand half-set of Ah and A arrays? (6) With all the necessary evidence at their disposal how often do members of the two groups reach the correct conclusion about the truth-value of a statement? (7) If the class to which the subject-term of a statement refers is empty, do subjects say that the statement is true, false or neither, and is there any difference between the groups in this respect? The answers to these questions serve to confirm the conclusions of the previous chapter, not only so far as the general superiority of the GR group in this area is concerned, but also with reference to more specific points. In the concluding section of the chapter some of the wider implications of the findings reported are considered.

4.1 Introduction Frequent references were made at the end of the previous chapter to a later experiment involving the same subjects in which an attempt was made to carry the investigation of the differences between the GR and PR groups one stage further. As the title of this chapter will make clear, I refer to the modified form of the card-turning task used in this second experiment as the 'four types of statement task' to distinguish it from the 'five types of statement task' described in the previous chapter. The number of statements involved has, in fact, nothing essential to do with the nature of the task; it simply provides us with a convenient label for the two tasks - which I shall refer to, for short, as the '5TS' and '4TS' tasks respectively.

The original intention in this later experiment was, of course,

to focus attention on the two types of statement, A and F, on which significant differences between the groups had appeared, in particular with a view to establishing whether or not these differences were to be understood in terms of the 'important' mistaken assumption that A and F statements imply the truth of their converses rather than being attributed to some 'unimportant' feature or features of the experimental task. The intention had been to use two other types of statement (Ah and Ad) as practice materials - these being chosen in preference to the fifth type of statement used in the 5TS task, E, because of the doubts we have seen there to be about the extent to which the difficulties of this type of statement are adequately reflected in the card-turning task. Preliminary trials with the 4TS test material made it clear that only one of the two main types of statement could be dealt with within the hour at the experimenter's disposal. Accordingly, F appears only in truncated form.

I have already mentioned one main purpose of the 4TS experiment: to establish whether the differences between the experimental groups in their success with the card-turning task as it relates to the A type of statement should be interpreted as reflecting a difference in the extent to which GR and PR subjects suppose that an A statement implies the truth of its converse. Other purposes were as follows: (1) to confirm the existence of differences between the groups in the adequacy with which they select cards for the A and F types of statement; (2) to throw light on the extent to which the responses of my subjects were stable over the period of twelve to eighteen months which elapsed between the 5TS and 4TS experiments; (3) to provide additional evidence on the question whether it is more difficult to recognise cards as exceptions to a rule when the first named character is on the reverse side (above, p. 104); (4) to study the extent to which the members of the two groups could learn to select the appropriate cards given information about the adequacy of previous selections; (5) to discover how subjects would handle the special case where there are no cards with the character first named (in the A type of statement).

It will be apparent that only the first two of the purposes described above are such as to provide us with the ^{of an explanation} elements of failure in a syllogistic reasoning task such as Section B of the Valentine test. The others relate to more general aspects of the card-turning task or of the reasoning processes of the subjects, but in widening the scope of the investigation in this way we are, I think, at the same time strengthening

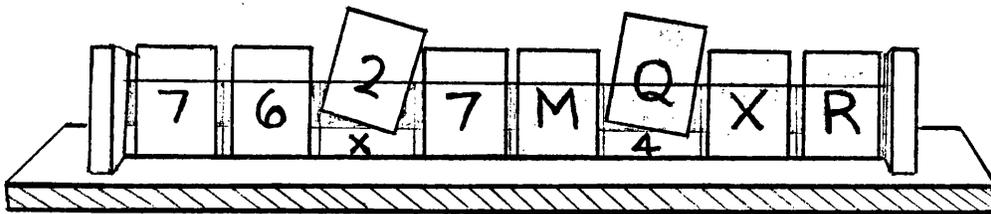
the interpretative basis on which our conclusions rest. How the various purposes are to be served in the 4TS task will become apparent in later sections when the details of the materials and procedure are presented.

4.2 The subjects Of the 44 subjects involved in the 5TS experiment 43 were able to return for the later task. The exception has already been referred to (p. 71); in the results presented below the number of pairs is, consequently, 21. The elapse of time between the two sessions varied from 12 to 18 months depending on the availability of subjects - some of whom were studying abroad when the others first began to report at my room for the second session. Needless to say, there was no systematic difference between the groups in this respect. In all cases memory for the events of the earlier session appeared to be extremely poor and, in any case, as I explained in the previous chapter, subjects were not told at any point in the 5TS task what the correct solutions were.

4.3 The 4TS materials For the 4TS task I prepared sets of actual cards and simplified the arrangements which obtained in the 5TS task (a) by using cards with a white side and a coloured side and (b) by using only letters and numbers for characters, there always being a letter on the coloured side of a card and a (single-digit) number on the white side. Subjects were faced, then, with an array of eight cards, the four on the left showing their white sides, the four on the right their coloured sides. The cards were held upright in transparent pockets on the back and towards the bottom of which the symbol on the reverse side of the card was written. Subjects could therefore discover what was on the back of a card simply by lifting it a small way out of its pocket, where it could be held by tilting it to one side. (See Figure 4.1.)

FIGURE 4.1

SAMPLE ARRAY OF CARDS AS USED IN THE 4TS TASK



This substitute for turning the card over was primarily a time-saving

device, but it did mean (a) that a check could readily be kept on the cards the subject had 'turned over' and (b) that once a card had been 'turned over' the subject could see the characters on both of its sides at the same time.

For reasons briefly indicated earlier (Section 4.1) fifteen sets of cards were prepared, two for Ah and Ad types of statement, nine for A and two for F. The function of the first four sets was originally intended to be to ensure that the nature of the task and of the materials was clear. The long series of nine was to have been the first of two, of which the two F sets were all that could be retained of the second in view of the unexpectedly long time some subjects took to complete the various tasks.

The special purpose of the long series was to see what would happen when a subject was told that he had made a wrong choice of cards to turn over and/or that he was mistaken in his conclusion about the truth-value of the statement relating to a particular set of cards. The series was carefully arranged to expose the subject to a variety of contingencies, some of them designed to prompt him, some of them to test the firmness of his understanding of the principles involved in the correct selection. It proved impossible to estimate in advance, even after a series of preliminary trials, how many sets of cards would be necessary in order to ensure that all subjects would eventually learn to make the correct selection: indeed since a large number never did learn to do this, this is something to which I still do not know the answer. At the other extreme, some subjects made the correct selection at the first attempt, and these had to be encouraged to complete the whole series of nine, at first by the promise, and subsequently by the reality, of more interesting cases towards the end of the series.

The fifteen sets of cards, together with the accompanying statements, are set out in Figure 4.2. I have followed Wason's practice of enclosing the character which appeared on the reverse of a card in brackets. Other features which call for explanation are the lower case 'x's and exclamation marks which appear below some of the cards. The first of these symbols indicates cards which are inconsistent with the truth of the accompanying statement; the second marks those cards which might be supposed to be inconsistent with the statement by a subject who supposes that the truth of the statement implies the truth of its converse.

FIGURE 4.2

THE SETS OF CARDS AND ACCOMPANYING STATEMENTS USED IN THE 4TS TASK

- Ah(a): 7(K) 6(X) 2(Z) 7(P) M(7) Q(7) X(8) R(4)
 x ! x x x !
 If a card has a 7 on one side it has an X on the other.
- Ah(b): 4(M) 4(Y) 3(S) 4(T) Y(4) X(8) M(2) R(4)
 x x x
 If a card has a 4 on one side it has a Y on the other.
- Ad(a): 8(A) 2(B) 7(C) 4(B) B(2) B(8) A(8) B(3)
 x* *
 Every card has either an 8 on one side or else a B on the other.
- Ad(b): 8(K) 2(K) 4(K) 2(K) K(4) K(7) M(2) N(2)
 *
 Every card has either a 2 on one side or else a K on the other.
- A(a): 2(T) 3(Q) 2(T) 4(K) T(2) T(2) S(4) T(2)
 All the cards with a 2 on one side have a T on the other.
- A(b): 4(S) 4(S) 3(T) 1(L) S(4) R(3) S(2) M(7)
 !
 All the cards with a 4 on one side have an S on the other.
- A(c): 7(P) 8(P) 9(S) 8(P) A(3) P(8) K(2) P(8)
 !
 All the cards with an 8 on one side have a P on the other.
- A(d): 3(X) 2(R) 1(S) 3(X) X(3) X(3) Z(3) Y(4)
 x
 All the cards with a 3 on one side have an X on the other.
- A(e): 2(B) 0(F) 0(B) 0(F) F(3) C(4) L(2) F(0)
 x !
 All the cards with a 0 on one side have an F on the other.
- A(f): 3(M) 4(N) 2(P) 4(K) K(7) M(7) K(7) K(6)
 ! x !
 All the cards with a 7 on one side have a K on the other.
- A(g): 5(E) 4(E) 4(R) 4(M) S(4) T(6) B(6) A(6)
 x x x
 All the cards with a 6 on one side have an E on the other.
- A(h): 3(A) 7(B) 1(C) 8(C) E(4) K(2) L(2) S(7)
 All the cards with a 5 on one side have an M on the other.
- A(i): 3(V) 4(R) 7(V) 8(S) V(9) V(8) V(9) V(8)
 ! ! ! !
 All the cards with a 9 on one side have a V on the other.
- F(a): 2(T) 3(Q) 2(T) 4(K) T(2) T(2) S(4) T(2)
 Only cards with a 2 on one side have a T on the other.
- F(b): 7(W) 7(W) 5(R) 3(L) W(7) X(2) W(6) P(4)
 x
 Only cards with a 7 on one side have a W on the other.

Needless to say, the statements accompanying sets of cards some of which have an 'x' below them are false, all others being true. I have mentioned on a previous occasion the ambiguity of the universal-disjunctive (Ad); those subjects who interpreted it in the exclusive sense would regard the cards asterisked as inconsistent with the truth of the accompanying statements.

It is perhaps necessary to say a little about the ways in which the above array of cards were expected to serve the purposes outlined at the beginning of this chapter. Attention has already been drawn to what is perhaps the most important of these - the inclusion of the cards marked in Figure 4.2 by exclamation marks to test the hypothesis that the failure of PR subjects to deal as successfully as GR subjects with A statements is to be attributed to a greater tendency on their parts to assume that statements of this type imply the truth of their converses. It will be noted that cards of this kind were also included in one of the sets relating to the other form of the universal affirmative, Ah.

The series of nine A sets of cards can be broken into two sub-groups, the first five and the remaining four. In the former group the problem is presented in an entirely straightforward form: succeeding sets help to draw the attention of the subject to the important features of the situation as opposed to testing the strength of his understanding, whereas the emphasis is in the opposite direction in the final four trials. A subject who is guilty of the confusion mentioned in the previous paragraph, for example, is likely to come to the wrong conclusion about the truth-value of the A(b) statement - and ought to learn from being told that his conclusion is wrong. Similarly, a person who fails to make the correct response so far as the right hand half-set is concerned is liable to find himself forced to revise his view about the truth-value of A(d) in the light of the seventh card of that set.

In the first of the sub-group of four, on the other hand, a subject whose understanding of the principles governing a correct selection is weak may find it impossible to restrict his choice to the single relevant card, there being no cards with the first named character on the sides of the cards facing him. In the following set there are no cards with either of the named characters on their exposed sides, and to succeed here the subject has to turn over all the cards in the right hand half-set - although, of course, these are all of the type which has been shown to

present most difficulty. When these cards are turned, as, of course, they eventually are by all subjects, it is clear that there are cards with the first named character in the set (so that the statement about the set is false).

In the following set the subject sooner or later discovers that there are no cards with the character first named: the subject term of the statement refers to a class of objects which is empty. This condition was included, not only because it appears to be a natural extension of the process which we have seen developing in earlier sets but also because it touches on a division of opinion among logicians as to the proper interpretation of the universal affirmative. Given my general interest in the way my subjects understand this, and other, types of statement, it seemed natural to ask the question whether the members of one group or the other favoured one of the two possible interpretations. Orthodox modern logic requires that we interpret the universal affirmative as a hypothetical to the effect that if there are any objects of a certain kind then they all possess such and such a property. Such a statement is true if there are no objects of the kind in question. Opposed to this interpretation (adopted by Quine (1952) for example) is the view of Strawson (1952) according to whom a person who makes a statement of the universal affirmative variety is understood to imply - in some sense of the word - that the class referred to in the subject term is not empty. If this proves not to be the case, one of the informal 'rules' governing the use of the statement has been broken and we should say, not that the statement is true but that it is neither true nor false.

As I have been implying, I thought it would be of some interest to discover which of these views my subjects favoured, and, more particularly, whether there was a difference between the groups in this respect. Finally, in the series of nine A statements, subjects were faced with a case in which none of the cards in the left hand half-set bore the character first named on their exposed sides while all of the cards in the right hand half-set showed the character second named. This situation is related to the one I have just discussed in the respect that a subject who takes the Quine view of the universal affirmative should say that he can tell that the last statement is true without turning any cards over, for there is no card which could count as an exception. A subject who takes a Strawson view of the situation, on the other hand, should say that all the cards in the right hand half-set must be turned so that he can

establish whether the precondition of truth - that the subject class should have members - is fulfilled. In fact it is and there remains only the question whether, given the characters on the two sides of the various cards, the statement is in fact true. To anyone with a clear grasp of the situation this is scarcely a question at all, for, as I have remarked, it is clear from the outset that none of the cards can be exceptions to the rule. Nevertheless, several subjects reached the wrong conclusion about this last statement because they regarded the cards marked with an exclamation mark in Figure 4.2 as exceptions. In other words, they supposed that 'All X's are Y's' implies that all Y's are X's: the confusion to which we have repeatedly returned in the course of this thesis asserts itself to the end.

4.4 Procedure Subjects were tested individually in my room. They were first of all reminded, in an informal way, of the terms of the problem and then introduced to the changes in materials and arrangements. As we have seen, these included the use of actual cards instead of diagrams, the simplification and clarification of the cards themselves - every card having a white side, with a letter, and a coloured side, with a number - and the method of turning the cards over. Subjects were then given a booklet to work from, the first page of which is reproduced, somewhat reduced, in Figure 4.3.

FIGURE 4.3

SAMPLE PAGE FROM 4TS BOOKLET (REDUCED)

In this set (a) if a card has a
7 on one side it has an X on the
other. True or false?

Cards to turn over:	7	6	2	7	M	Q	X	P
Statement true	<input type="checkbox"/>							
Statement false	<input type="checkbox"/>							
Critical card(s):	7	6	2	7	M	Q	X	P
ALL CARDS TURNED OVER								
Statement true	<input type="checkbox"/>							
Statement false	<input type="checkbox"/>							
Critical card(s):	7	6	2	7	M	Q	X	P
Cards to turn over: (final view)	7	6	2	7	M	Q	X	P

The more detailed instructions which follow were then read to the subject, every effort being made to ensure that all important points were grasped.

Instructions At the top of each page of this booklet you will find a statement about the set of cards which is displayed in the stand in front of you. As on the previous occasion, your task is to say which of these cards you need to turn over in order to establish whether the statement is true or false. You are to indicate your choice of cards by putting crosses in the corresponding places on the page opposite the words 'Cards to turn over'.

After you have done this, you are to turn over the cards you have crossed. (Here followed a reminder about the way in which cards were to be 'turned'.) When you have done this, you should be in a position to say whether the statement is true or false. You are to indicate which by placing a cross in the appropriate box opposite 'Statement true/false'.

If the statement is false, this must be because at least one of the cards you have turned over is 'out of line', runs counter to the rule. These are the 'critical cards' and once again you are to mark them with a cross in the appropriate place(s).

Next you are to turn over any cards which you have not so far turned over. The point of this is to ensure that all subjects are in possession of the same information at this point and that all are exposed to the same 'prompts' and 'trips'. You are then to say whether the new cards you are exposed to change your mind about the truth or falsity of the statement. Once again, place a tick in the appropriate box.

If you have changed to 'False', then you will have 'critical cards' for the first time in that set and these should be indicated in the appropriate way, by putting crosses in the relevant boxes, opposite the second 'critical card(s)'. Even if you do not change your mind at this stage, if your view was that the statement was false, you may find that there are some critical cards which you had not suspected. You should indicate these, as well as the others, in the boxes opposite the second 'critical card(s)'.

If you do not change your mind either about the truth or falsity of the statement or about the cards which are critical, you will have crosses in the same boxes in the fourth and fifth lines as you had in the previous two.

Finally, you may in the course of all this have changed your view as to the cards which need to be turned over in order to prove the statement true or false. If you have changed your view, then you have a chance to indicate this in the last line on the page. If there is no change in your choice of cards, then, of course, you will place crosses in the same boxes in the last line as you did in the first.

When you have done this, I shall tell you whether the statement really is true or not, and also whether your choice of cards is the correct one or not. If I say it is not correct, this may mean a variety of things - that you have turned over cards you need not have turned, that you have not turned some cards you should have turned or, of course, both. I shall not tell you which of these alternative possibilities obtains in your case.

The whole situation is supposed to be one in which subjects who make the wrong choice to begin with learn to make the correct choice. Now, of course, you may get it right to begin with - and in this case you are asked to be patient and careful not to make accidental errors. If, on the other hand, you get it wrong, there can be no guarantee that you will discover the correct solution even in the long series of nine; on the contrary, a large proportion of students do not get it right by the end of the series.

We begin with two sets of cards for each of two types of statement. The function of these sets is to ensure that the procedure is clear. If you have any doubts on this score, please do not hesitate to say so. Then there is the long series of nine sets of cards all involving a third type of statement, and, finally, a further two sets of cards relating to a fourth type of statement. Between the long series and the final two sets of cards I shall ask you to complete Form B of the personality questionnaire of which you did Form A last time. This will give you a break after the long series.

You will not be timed. Any questions?

In the event subjects had difficulty with only two aspects of the procedure as outlined above: they had to be reminded that there were critical cards only if the statement was false (or believed to be); and they had to be reminded to repeat their earlier responses in cases where no change in their views occurred after all the cards were turned over. Since neither of these affects the validity of responses to the Ah and Ad statement-types, and since any warming-up effects that there were would be the same for members of the two groups (no inter-statement comparisons being made) I have treated performances on these first two statement-types as comparable in validity with responses to the A and F series.

So far as these latter are concerned, it is important to note that all subjects who had not discovered the correct choice of cards by the end of the long series were told what this was, emphasis being laid on the point that that other selections would generally be called for in the case of other statement-types. Since all subjects did all the sets of cards in the same order, performance on the F statement-type might be supposed to be affected by 'transfer' of a positive or negative kind: negative if subjects made the same choice of cards as they had discovered - or had been told - was correct in the case of A statements, positive if their understanding of the nature of the task had been increased by one or other of these means. To some extent it was hoped that any such effect would be reduced by the fact that all subjects completed Form B of the Eysenck Personality Inventory between the ninth set of A and the first set of F, the completed questionnaire also being scored and the results

discussed. (This aspect of the present research is the subject of a section of the following chapter.) In any case, it seemed unrealistic to suppose - and events bore this out - that it would be possible to take subjects through this extended 'learning procedure' and not be faced with the demand, at least from some subjects who failed to discover the correct solution, that I explain what was wrong with the solution they had offered. It seemed unwise to risk finding myself faced with a choice between losing my subject's cooperation and introducing ad hoc what would clearly be an important difference between subjects.

4.5 Results: (a) Stability of response from 5TS to 4TS In its simplest terms the question to be answered in this section is whether subjects responded to the various statement-types in substantially the same way in the two experimental sessions, occurring, as they did, with an elapse of twelve to eighteen months between them. The question is of interest quite generally as shedding light on an aspect of the card-turning task, but also more specifically as indicating the extent to which the behavioural patterns, and the mediating responses which are assumed to lie behind them, represent real or continuing differences between the members of the two groups.

In view of the changes in the selection of cards which are likely to have been produced by the 'learning' component of the 4TS task, it seems clear that any attempt to estimate stability of response must consider only the first selection of cards for each of the four statement-types in the 4TS task. We can compare this selection with the selection made either in the first or in the fifth trial of the 5TS task, depending on whether we are interested in the extent to which the first reaction to the statement-types was the same on both occasions or in the extent to which changes occurred in the interval between the last trial of the 5TS, and the first trial of the 4TS, task. Finally, inasmuch as our interest is in the stability of the responses over a relatively long period of time, it is clearly important to compare the changes from 5TS to 4TS with the degree of change occurring within the earlier task itself (in the course of which, as we have seen, no information was given to the subject about the correctness or otherwise of his responses).

A measure of correlation is, of course, what we need. Unfortunately, however, I have been unable to find any which will cope with the kind of data in question. The difficulty arises from the fact that, as

we saw in discussing the 5TS task, there are important qualitative differences in response to the card-turning task which it is impossible to represent unequivocally in numerical terms. There are four types of card, each of which may be treated appropriately or otherwise in relation to one particular type of statement, and while it is plausible to maintain, as I have done, that a subject who treats all four types correctly has responded more adequately than a subject who succeeds with only three - and so on through the cases where two, one and none of the types of cards receive the correct treatment - in a scoring system which awards 4, 3, 2, 1 and 0 points respectively in these five cases, only the first and last of these scores are unambiguous in the sense that it is possible to say that a subject who has the same score on different occasions has made the same response on each: a subject might in principle make a different response on each of five trials and still score two points on each.

In the absence of a more appropriate measure of correlation, then, I simply counted the number of changes made by each subject in his response to each of the four types of statement occurring in the 4TS task - from 5TS trial 1 to 5TS trial 5, from 5TS trial 1 to 4TS trial 1, and from 5TS trial 5 to 4TS trial 1. The results are presented in a variety of ways in Tables 4.1 to 4.4. A 'change' in these tables is a change in a subject's treatment of one of the four types of card involved in any statement-type - from turning all the cards of that type to leaving them all unturned, or vice versa, or from either of these 'complete' responses to an 'incomplete' response (in which, it will be recalled, some cards of a type are turned and some are not) or, again, vice versa. The maximum possible number of changes in any one statement-type from one trial to another is, of course, four.

Table 4.1 shows that for both groups the amount of change from the last trial of 5TS to the first trial of 4TS is very similar to the amount of change from the first to the last trial of the earlier experiment. As between the first trials of the two experiments the amount of change is somewhat greater. Whether this is to be regarded as a large or small amount of change is not entirely clear. In theory, subjects might have made as many as sixteen changes from any one trial to any other and this might seem to suggest that the actual number of changes made is small. On the other hand, as we saw in the previous chapter, certain components of the response to a particular statement-type are almost universal (turning over cards with the character first named in Ah and A statements,

TABLE 4.1

MEANS, STANDARD DEVIATIONS AND RANGES OF 'CHANGE' SCORES FROM 5TS(1) - 5TS(5), 5TS(1) - 4TS(1) AND 5TS(5) - 4TS(1): ALL FOUR STATEMENT-TYPES

		From 5TS(1) to 5TS(5)	From 5TS(1) to 4TS(1)	From 5TS(5) to 4TS(1)
GR group	Mean	3.7	5.5	4.0
	S.D.	2.4	2.9	3.1
	Range	0 - 10	1 - 12	0 - 13
PR group	Mean	3.9	5.0	3.8
	S.D.	2.7	2.4	1.7
	Range	0 - 10	1 - 9	0 - 6

for example) and little change is to be expected in such cases. At the same time, as Table 4.1 makes clear, some subjects did make a large number of changes from one trial to another and more information about the incidence of changes of different dimensions in different statement-types is presented in Table 4.2.

TABLE 4.2

NUMBER OF SUBJECTS MAKING DIFFERENT NUMBERS OF CHANGES FROM 5TS(1) - 5TS(5), 5TS(1) - 4TS(1) AND 5TS(5) - 4TS(1) IN DIFFERENT STATEMENTS

No. of changes	5TS(1) - 5TS(5)				5TS(1) - 4TS(1)				5TS(5) - 4TS(1)			
	Ah	Ad	A	F	Ah	Ad	A	F	Ah	Ad	A	F
0	9	15	9	7	8	12	4	4	9	15	7	7
1	8	-	8	5	5	-	12	3	10	-	8	6
GR group	2	3	4	4	4	4	3	8	-	3	4	2
3	1	-	-	4	4	1	2	4	2	1	2	6
4	-	2	-	1	-	4	-	2	-	2	-	-
Total no. changes	17	16	16	29	25	27	24	30	16	17	22	28
Ave. over group	.8	.8	.8	1.3	1.2	1.3	1.1	1.9	.8	.8	1.0	1.3
Average overall	.92				1.36				.98			
0	12	11	10	6	8	8	9	1	12	10	13	6
1	5	2	4	5	9	3	7	8	5	1	5	7
PR group	2	4	7	5	3	5	2	8	4	6	2	3
3	-	-	2	1	1	2	3	4	-	2	1	5
4	-	1	1	1	-	3	-	-	-	2	-	-
Total no. changes	13	20	20	28	18	31	20	36	13	27	12	28
Ave. over group	.6	1.0	1.0	1.3	.9	1.5	1.0	1.7	.6	1.3	.6	1.3
Average overall	.96				1.25				.95			

Table 4.2 enables one to see, not only how often the larger numbers of changes within statement-types occurred, but also, of course,

whether changes were more likely to occur in one statement-type rather than another and whether the two groups differed in this respect. So far as the last of these questions is concerned, the differences in 'change scores' between the two groups were not significant at the 5 per cent level on the Wilcoxon Signed Ranks Test, whether the scores on the four statement-types were ^{taken} separately or together. As to the other points, the largest number of changes occurred in connexion with the F type of statement for all three comparisons and for both groups. This is perhaps to be expected in view of the striking diversity of response to this statement-type noted in the previous chapter and again with reference to Table 4.5 below. And finally, in connexion with Table 4.2, it was clearly rare for a subject to make changes in his treatment of all four types of cards, somewhat less rare for him to change his treatment of three of the four types - though the comparison between the first and last trials of the 5TS experiment provides something of an exception in this respect - and increasingly common to change his treatment of two, one and none - though there is, once again, an apparent exception to this rule in the PR group's performance on the first 4TS trial as compared with its performance on the first trial of the previous experiment.

Tables 4.1 and 4.2 show that the amount of change from the first to last trials in the 5TS experiment and from the last trial of 5TS to the first of 4TS is smaller than from the first of 5TS to the first of 4TS. A question naturally arises about the extent to which the changes from 5TS(1) to 5TS(5) and from 5TS(5) to 4TS(1) can be supposed to have been changes in a single direction - and, in particular, about the extent to which they represent a progressive improvement in subjects' responses to the various statement-types from 5TS(1) to 4TS(1). Tables 4.3 and 4.4 are intended to provide an answer to this question.

Table 4.3 presents the numbers of subjects who made the appropriate response to each of the four types of card in each of the four statement-types on the three trials presently under consideration. Clearly, these numbers should increase from ^{one} trial to the next if a gradual improvement did occur. I have asterisked the cases in which the opposite occurred. There are, in fact, only two such cases out of the twenty-eight in which a change in rate of success occurred for the GR group; the corresponding figures for the PR group are six out of twenty-six. On the whole these results are fairly strong evidence for the view that there was a steady improvement in the performance of subjects over the three

TABLE 4.3

NUMBER OF SUBJECTS GIVING THE APPROPRIATE TREATMENT TO THE FOUR TYPES OF CARDS OVER THE THREE TRIALS, 5TS(1), 5TS(5) AND 4TS(1)

Type of Card: ¹		S t a t e m e n t T y p e															
		Ah				Ad				A				F			
		P	\bar{P}	Q	\bar{Q}	P	\bar{P}	Q	\bar{Q}	P	\bar{P}	Q	\bar{Q}	P	\bar{P}	Q	\bar{Q}
GR group	5TS(1)	16	17	9	3	12	15	11	15	19	19	8	5	7	11	12	11
	5TS(5)	19	20	13	6	16	17	16	16	19	20	15	7	9	16	16	17
	4TS(1)	21	20	13	10	18	20	18	20	21	20	12*	12	15	14*	19	18
PR group	5TS(1)	17	18	9	2	11	11	10	9	15	17	9	2	3	5	13	16
	5TS(5)	19	19	9	4	13	13	14	13	20	19	8*	2	5	16	15	11*
	4TS(1)	21	19	7*	1*	18	16	18	18	20	19	7*	3	7	16	13*	15

* Asterisks denote trials on which a lower rate of success is achieved on that trial than in the preceding one.

trials. Hardly surprisingly, improvement is on the whole less obvious from the last trial of 5TS to the first of 4TS than from the first to the last of the former.

The same point about the overall tendency towards an improved performance over the three trials can be made by reference to the fact that on all 16 components the GR group was more successful on the first trial of the later experiment than on the first trial of the earlier one, the corresponding figure for the PR group being 11 out of 16, with one component in which the group's performance on the later was the same as on the earlier one. One other aspect of the data presented in Table 4.3 is that the GR group made advances, over the three trials, at just those points where one would expect an advance, viz., where their original performance was poorest. No such tendency is apparent in the case of the PR group.

Table 4.4 attempts to clarify the extent to which change over the three trials represented improvement. It will be clear from this Table that change was more likely to represent net improvement (a) in the case of GR subjects and (b) when the first and last trials of the earlier

¹ For typographical convenience in this table I have used Wason's symbols to designate the four types of cards: 'P' for cards with the character first named, ' \bar{P} ' for the other cards in the left hand half-set, 'Q' for cards with the character named second, and ' \bar{Q} ' for the other cards in the right hand half-set.

TABLE 4.4

PROPORTION OF CHANGES OVER THE THREE TRIALS REPRESENTING NET IMPROVEMENT

		<u>Statement</u>		<u>Type</u>	
		Ah	Ad	A	F
GR group	5TS(1)- 5TS(5)	.76	.75	.63	.58
	5TS(5)- 4TS(1)	.38	.64	.18	.28
PR group	5TS(1)- 5TS(5)	.38	.60	.30	.35
	5TS(5)- 4TS(1)	-0.23*	.62	.00	.14

*The minus sign represents net deterioration

experiment are compared - the second of these results, as already noted, being just what one would expect, given that the interval between these trials was a relatively short one and filled with practice (albeit without knowledge of results) of the task in hand. One other point from Table.4.4 is that the proportion of change representing improvement is considerable, for the PR group, only in the case of the universal-disjunctive (Ad) type of statement.

Finally, in connexion with the stability of response to the card-turning task over the period of 12 to 18 months, it seemed useful to present data bearing on the extent to which the same types of response to the various types of statement were preferred in the three trials, and this is done in Table 4.5. This table may be compared with Table 3.12 above where the preferences of my groups in the 5TS task are compared with the preferences of Wason's various groups. The classification of responses proceeds on the same principle - the extent to which each of the four types of cards in turn is treated in the appropriate way or not, so that a '++++' response is one in which all four types are given the appropriate treatment, a '+++-' response is one in which the fourth type (\bar{Q}) is given the wrong treatment, and so on.

This approach to the question of stability of response is, of course, rather different, at least in principle, from the one adopted at the beginning of this section, since it relates to the extent to which

TABLE 4.5

FREQUENCY OF DIFFERENT TYPES OF RESPONSE IN 5TS(1), 5TS(5) AND 4TS(1)

		S t a t e m e n t T y p e s											
		A h			A d			A			F		
		5TS (1)	5TS (5)	4TS (1)	5TS (1)	5TS (5)	4TS (1)	5TS (1)	5TS (5)	4TS (1)	5TS (1)	5TS (5)	4TS (1)
GR group	++++	2	3	7	11	16	17	1	6	7	3	6	12
	+++-	7	9	6				7	8	5			
	++-+	1	3	2				3	1	4		1	
	+--+										1	1	3
	-+++										2	5	
	++--	5	4	5				7	4	4	2	1	
	+---			1				1		1			
	----										3	3	3
	+--+						1						
	-+-+				4		3				1		
	-++-		1						1		3	1	1
	+---	1			1						1		
	-+--	2				1		1				2	1
	---+										1	1	
	----	3	1		5	4		1	1		4		1
PR group	++++		3		7	9	5		1	1		3	4
	+++-	9	6	7	1			6	6	6			
	++-+		1						1	1	1	1	1
	+--+						2				2		
	-+++											2	4
	++--	6	8	12				7	11	11		1	2
	+---	1		1				1					
	----										8	3	2
	+--+				2	4	1		1				
	-+-+	1			2	4	1				2	1	2
	-++-							2			2	6	2
	+---	1	1	1	1			1		1			
	-+--	2	1		1			2	1			2	1
	---+					1		1			1	1	1
	----	1	1		7	3	2			1	3	1	2

the responses of the group as a whole were the same, and not to the extent to which individuals tended to make the same response: in principle, at least, a given type of response might be made equally often by the group as a whole on subsequent trials without any one member of the group ever making it on successive occasions. In this respect the information presented in Table 4.5 is of less immediate relevance to the question of the reliability of the card-turning task as an index of the subjects' grasp of the meaning of the various statement-types than the information presented in earlier tables of this section; at the same time, it does provide

evidence relating to the more general question whether, in this type of task, the same responses are favoured by the same group on different occasions as we have seen, in the previous chapter, they tend to be when the responses of different groups are considered. Inspection of the table suggests the following comments.

In general, the responses preferred in 4TS(1) are the same as the ones preferred in the 5TS trials, matching responses¹ in all four types of statement and half-matching responses¹ in Ah, A and F statements being particularly favoured. Apparent exceptions are the absence of all-minus responses in Ad for the GR group, the increased incidence of all-correct responses in F for the same group, and an increase in the number of ++-- responses in Ah for the PR group. The first two exceptions are probably best seen as reflecting the general improvement in adequacy of response which we have noted in the case of the GR group over the three trials. It is less plausible to attribute the third exception to the same source: what it does appear to reflect is a reduction in the number of different types of response considered appropriate in the Ah case by the PR group - though I suppose this might be said to be in itself evidence of improvement of a rather general kind. Meanwhile, the great diversity of response to the F type of statement noted in connexion with Table 3.12 is continued in 4TS(1), although perhaps somewhat reduced in the case of the PR group.

In summary of this section, then, the responses of both groups to the four types of statement included in the later experiment are generally rather similar to the responses made on the previous occasion, there being, however, fairly clear evidence, particularly in the case of the GR group, of a slight improvement in response from the one experiment to the other.

4.6 Results: (b) Evidence of learning over subsequent trials

Having seen that the responses of the two groups were relatively stable from 5TS(5) to 4TS(1) we might expect that the differences between the two groups reported in relation to the 5TS experiment would be confirmed in the results of the later one. However, all selections of cards in the 4TS experiment after the first were liable to be affected by the subject's experiences in turning over all the cards in the set, and all selections

1 These terms were introduced in the previous chapter: a matching response is one in which the subject turns over only the cards with named characters, a half-matching response one in which he turns only cards with the character first named.

after the second by the knowledge of the correctness or otherwise of his choice. It seems appropriate at this point, therefore, to present evidence about the extent to which learning seems to have occurred as a result of these particular features of the 4TS task. Figure 4.4 shows, in graphical form, the way in which the scores of the two groups changed over the four selections of Ah, Ad and F, and the first ten or twelve selections of A, the scores for the two half-sets in each statement-type being presented separately.

Three points call for comment in this connexion: the kind of scoring system employed, the presentation of only 10 or 12 selections in the case of A, and the reduction of N to 20 in the case of F. So far as the first of these is concerned, the method of scoring employed throughout the discussion of the 4TS experiment is the third of those described in the previous chapter - where a subject gains a point for each type of card (and not, as in the fourth method, for each particular card) correctly dealt with. On this scoring system a person is as heavily penalised for an incomplete response as for a complete, but erroneous, one, and it is to be preferred to the fourth system, therefore, only if the number of subjects making incomplete responses for reasons irrelevant to an evaluation of their performance on this task is small. Such reasons would include perceptual failures and failures due to lapses of attention. By their nature such failures could be supposed to occur only infrequently in any one subject's performance - so that we may state the conditions in which the third method of assessment is the appropriate one as those in which only a small number of subjects make infrequent incomplete responses. As we saw in the previous chapter, these conditions did not hold in the 5TS experiment; in the later experiment, on the other hand, as Table 4.6 will make clear, this was no longer the case.¹

TABLE 4.6

SUBJECTS MAKING DIFFERENT NUMBERS OF 'INCOMPLETE' RESPONSES IN 4TS

	<u>Number of Incomplete Responses</u>											
	0	1	2	3	4	5	6	7	8	9	10	11
GR group	18	2	-	1	-	-	-	-	-	-	-	-
PR group	12	3	-	2	-	-	-	2	-	1	-	1

¹ Inspection of the results presented in Table 4.6 suggests a difference

As to the second point, it may be recalled that both categories of card appeared in the left-hand half-set only in the first five sets of A cards and in the right-hand half-set only in the first^{six}/. Hence the absence, in Figure 4.4, of scores for selections beyond the tenth and twelfth respectively, there being, of course, in every case two selections per set of cards, one at the beginning and one at the end of the relevant trial. Finally, one PR subject found the 4TS task as a whole so distasteful (no doubt because of her complete lack of success) that she could not be prevailed upon to proceed beyond the first two selections of F. I have therefore been obliged to leave her, and her GR counterpart's, performance out of account in the case of F - though her poor performance and her counterpart's success make it certain that such differences as exist between the two groups will be reduced by this procedure.

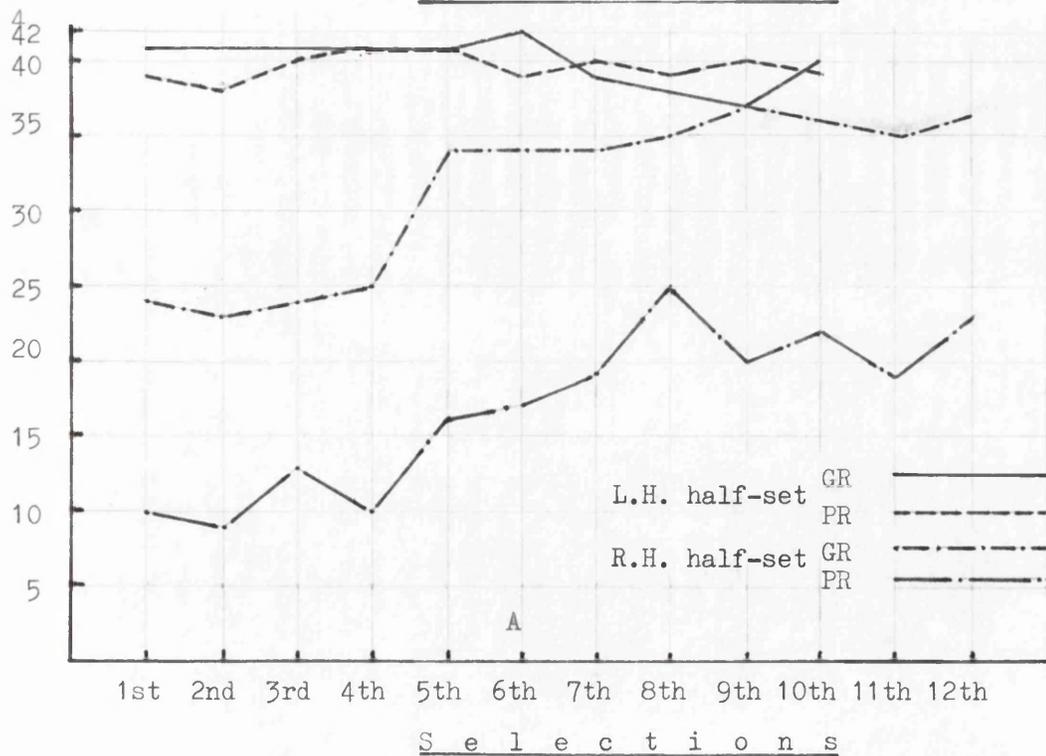
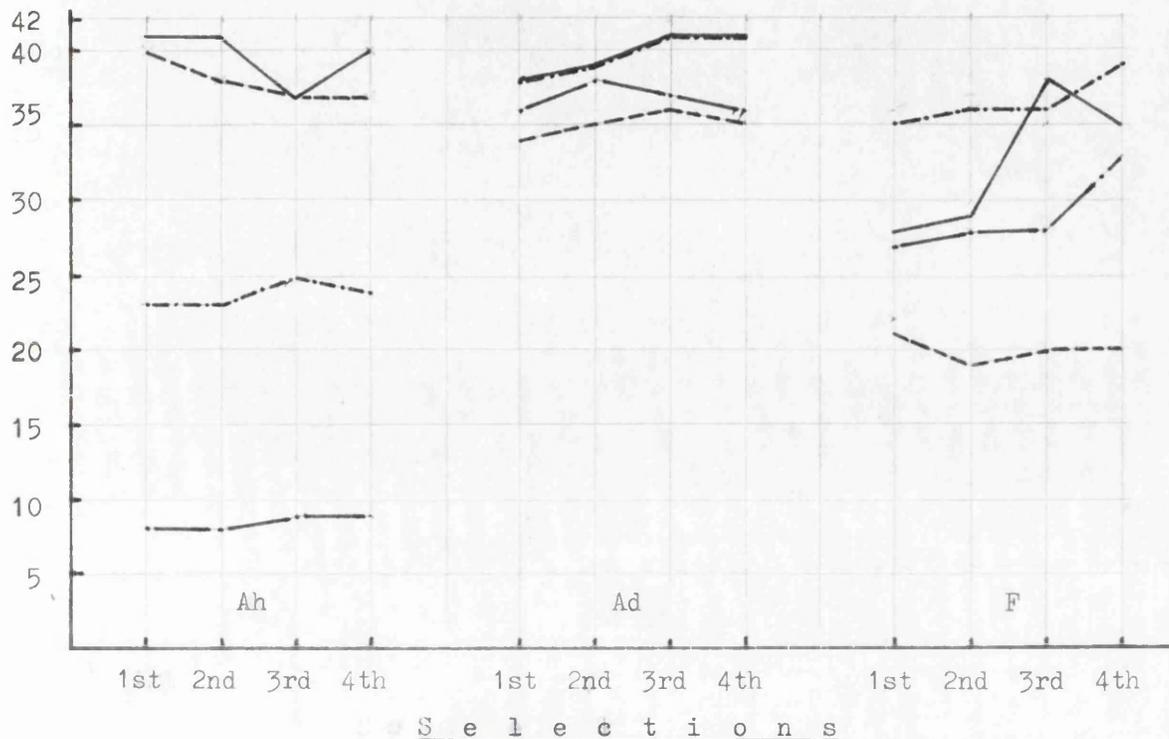
Wason (1968, 1969) has already shown that one kind of error made in response to the universal affirmative (whether in hypothetical or categorical form) is rather resistant to various 'therapeutic' measures intended to remove it. Since this error is also a very important one in the responses of my own subjects (it consists, essentially, in failing to see the relevance of \bar{Q} cards - cards in the right hand half-set which do not have the character named second) and since the 'therapy' represented by the procedure adopted in the 4TS experiment is minimal in character (disclosing the truth-value of the statement and the correctness or otherwise of the subject's selection) little evidence of learning might be expected.

So far as the Ah, Ad and F types of statement are concerned, it seems from Figure 4.4 that such learning as took place occurred in the PR group's treatment of the right hand half-set of F, in the GR group's treatment of the left hand half-set of the same statement-type, and less certainly, in the same group's treatment of the other half-set of this statement type. In view of the very limited opportunity which subjects had to make changes in their responses to these statement-types the most one can conclude is that the mistakes subjects were making were not of the kind which they could correct as soon as they were told that they were making a mistake. As for the F case, the fact that improvements do seem to have been

between the groups in the extent to which they were likely to make incomplete responses. However, a dichotomous division of the groups into those who did and those who did not make such responses gives a χ^2 (corrected for continuity) of 2.93, which is not significant at the 5 per cent level.

FIGURE 4.4

GROUP SCORES IN SUCCESSIVE SELECTIONS IN THE FOUR TYPES OF STATEMENT
 (HALF-SET SCORES PRESENTED SEPARATELY; MAXIMA: F = 40, OTHERS = 42)



made here must presumably be explained, not in terms of any special characteristics of this statement-type, but in terms of its position in the experiment as a whole. To be specific, the learning which seems to have occurred over the successive trials of the A type of statement may be supposed to have produced an improvement in subjects' ability to cope with problems of this type in general, a 'learning set', as it has been called, (Harlow, 1949) which facilitates the acquisition of adequate responses to subsequent problems of the same general kind.

In the case of the universal affirmative in its categorical form (A) there is once again little evidence of learning over the first four trials. This is hardly surprising so far as the left hand half-set is concerned, for neither group has much room for improvement. In the right-hand half-set the PR group shows a momentary improvement after the second selection which may presumably be attributed to the experimenter's comments on the adequacy of the subjects' responses so far. The same effect seems to be apparent after the fourth selection (at the end of the second trial). The gradual improvement in the PR group's performance on this half-set shows only one other spurt - between the first and second selections of cards in the fourth trial, where the first false statement, and the first critical card of this statement-type, is encountered. The gain is not consolidated, however.

The progress of the GR group in dealing with this half-set is also gradual - with the notable exception of the steep rise in score after the fourth selection, when, of course, subjects would be responding to the experimenter's second verdict on the adequacy of the solutions offered. Since, as we have seen, this comment consisted solely of a remark to the effect that the subject's selection of cards was right or wrong (and that the statement was true), it is necessary to assume, I think, that GR subjects had rapidly come to identify the probable locus of any errors they were making. Occasionally, a subject made this achievement explicit by asking the question (not, of course, answered), "Does 'All the X cards are Y cards' mean that all the Y cards are X cards?"

The gradualness of the improvement which we have seen to be represented, on the whole, in the curves of Figure 4.4 suggests that such learning as took place was of the trial-and-error variety (Thorndike, 1911). Subjects had been asked not to adopt the mechanical exploration of all possible selections of cards but to make changes in their selections only when they thought they could see why a different selection might be the

correct one. It seemed clear from the observation of the behaviour of individual subjects that a conscientious attempt was made to obey this instruction. At the same time, it seems likely that an uncertain amount of trial and error did take place - uncertain because of the difficulty, even for subjects themselves, of knowing in exactly what circumstances subjects could claim to 'see' why a change was justified. In a few cases, by contrast, something which could only be called 'insight' did seem to occur: the subject remarked, usually with an expression of relief and pleasure that he or she could now see what had been wrong with previous selections - and proceed to demonstrate this by making the correct response even in the later trials in the series where an adequate grasp of principle seemed to be a precondition of success. (On the importance of this aspect of problem-solving behaviour which deserves to be called 'insightful' see Woodworth and Schlosberg, 1954.) But finally, the curves of Figure 4.4 not only fail to reflect the differences between individuals who achieved a secure grasp of the principle underlying an adequate solution and those whose progress was more gradual and less certain, but also the difference between subjects of either of these categories and those who made no progress at all, and who revealed, not only by their continuing failure to make the appropriate changes in their selections but also by their remarks, that they had absolutely no idea how their responses should be changed to make them more satisfactory.

It may be of some interest to report, at this point, on one kind of failure to profit from the experiences provided by the 4TS task. It will be recalled that subjects were asked to say, when they had turned over all the cards in a set, which of them were 'critical' - which were incompatible with the truth of the relevant statement. They were then asked, in the light of this, to say which cards they ought to have turned over in order to establish the truth-value of the statement. A rather primitive type of response is to say only those cards which were actually critical - leaving out of account other cards belonging to the same category: for example, saying cards 5 and 6 in Ah(a) but not card 8, or card 8 in the following set but not cards 6 and 7. Such a response is primitive, I think, because the subject apparently fails to realise that what was in fact true of the cards he has marked might also have been true of any other cards of the same type. A failure of this type appears to be a failure to think in hypothetical terms, and, as such, a failure to think at the level of what Piaget (1950) calls 'formal operations'. (Wason has made a similar point in his 1968 paper.)

But if the failure just described is a primitive one, what is one to say of subjects who do not even draw the conclusion that cards which are actually incompatible with the truth of the statement in question should have been turned? Taking all subjects and all statement-types together there were respectively 19 and 37 occasions on which members of the GR and PR groups marked at least one card as critical and then failed to conclude that it ought to have been turned over. There were eight false statements in all, so that the total number of occasions on which such a failure could have occurred is 168 for each group. It is most unlikely that purely perceptual failures would have played an important role in this connexion for the arrangement of the information on the page from which subjects were working would be likely, if anything, to encourage a mechanical endorsement of cards ticked as critical. 13 of the 19 GR cases and 23 of the 37 PR cases occurred in the two Ah sets, both the relevant statements being, of course, false. Since these were the first two sets in the experiment, the failure under consideration may perhaps be attributed, in these cases, to the novelty of the situation - though it is not clear whether we should say that the failure is one which the subject learns very quickly to avoid, or whether it should be attributed to the non-specific factors in a problem-solving situation with which a 'warming-up' period is designed to cope. In any case, 6 PR and 5 GR subjects made a mistake of this kind on at least one subsequent occasion. Given the very high level of ability of my subjects and the elementary nature of the mistake, even this relatively small incidence of failure must be regarded as somewhat surprising.

4.7 Results: (c) The relative success of the two groups Figure 4.4 leads us to expect significant differences between the groups in performance on the right hand half-sets of Ah and A and (less certainly) on the left hand half-set of F. There appears to be no significant difference between the groups on the remaining statement-type, Ad, the universal disjunctive. Since the curves of the two groups are more or less parallel in Ah, Ad and A, it appears to be in order to sum the scores on successive selections and base our conclusions about differences between the groups on these. In the case of the left-hand half-set of F, where the curves for the two groups diverge rather sharply from the parallel, such a procedure would be justified, for present purposes, only on the assumption that differences in the extent to which the two groups of subjects understand the task is reflected in the speed with which they learn to make the correct selection of cards. This is not, in my view, an entirely improb-

able assumption; on the other hand, I suggested, in the previous section, that another factor, viz., the ability to benefit from previous learning experiences in the same kind of task, may be operative in the case of F and I have therefore tested the differences between the groups on F (both half-sets together and the lefthand half-set by itself) on separate selections as well as on all four taken together. Table 4.7 presents the mean scores of the two groups per selection of cards, there having been, of course, two selections for every set of cards.

TABLE 4.7

MEAN SCORE PER SELECTION OF THE TWO GROUPS ON THE FOUR TYPES OF STATEMENT

	Ah			Ad			A		
	L.H.	R.H.	Both	L.H.	R.H.	Both	L.H.	R.H.	Both
GR group	39.8	23.8	63.5	39.8	39.8	79.5	40.1	30.6	70.7
PR group	38.0	8.5	46.5	35.0	36.8	71.8	39.6	16.1	55.7
Difference	1.8	15.3	17.0	4.8	3.0	7.7	0.5	14.5	15.0
N subjects	21	21	21	21	21	21	21	21	21
Wilcoxon T	104.0	18.0	23.0	89.5	164.0	91.0	91.0	14.0	23.5
N ranks	21	21	20	21	21	21	20	21	20
p (2-tailed)	N.S.	<.01	<.01	N.S.	N.S.	N.S.	N.S.	<.01	<.01

	F (4 selections)			F							
	L.H.	R.H.	Both	1st		2nd		3rd		4th	
L.H.				Both	L.H.	Both	L.H.	Both	L.H.	Both	
	32.5	36.3	68.8	28.0	31.5	29.0	32.5	38.0	36.5	35.0	37.0
	20.0	29.0	49.0	21.0	24.0	19.0	23.5	20.0	24.0	20.0	26.5
	12.5	7.3	19.8	7.0	7.5	10.0	9.0	18.0	12.5	15.0	10.5
	20	20	20	20	20	20	20	20	20	20	20
	37.0	47.0	34.0	70.5	45.5	56.5	48.5	20.5	21.0	32.5	37.5
	20	19	20	20	19	20	20	20	20	20	19
	<.01	N.S.	<.01	N.S.	<.05	N.S.	<.05	<.01	<.01	<.01	<.01

It will be clear from an inspection of the above Table that the differences between the groups which were found to be significant in the 5TS experiment (on A and F as a whole, on the right hand half-set of Ah and A and on the left hand half-set of F) are also significant in the 4TS experiment, the level of significance being in every case higher in the later of the two. In addition to these the difference between the groups on Ah when the two half-sets are taken together is highly significant in 4TS whereas it failed to reach significance in 5TS. The later outcome is in fact the one which we should expect in view of the logical equivalence of Ah and A - and, as I have suggested in the previous chapter, the psychological equivalence of these statement-types as judged by the similarity

of response elicited by them. The fact that this difference between the groups failed to reach significant proportions in the earlier task must, therefore, be regarded as something of an anomaly. Finally, the results of the 5TS experiment are also confirmed in the case of the remaining statement-type, Ad, where the GR group once again did better than the PR group though not significantly so.

In making this summary of the results presented in Table 4.7 I have not thought it necessary to make a distinction between the outcome for F when all four selections are taken together and that obtained when they are considered individually. It is true that the difference between the groups on the left hand half-set alone is not significant when the first two selections are taken singly, but I think this may be attributed partly to a loss in the power of the Wilcoxon statistic when the range of possible scores is restricted, as it is in these cases, to three (0 - 2). In any case, in view of the uncertain effect on subjects of doing the F task after the long series of trials on A a certain amount of caution seems to be justified in interpreting these results as running counter to those obtained in 5TS.

4.8 Results (d) Interpretation of the difference between the groups on A and Ah The results reported in the previous section are important in the respect that they confirm the general conclusions drawn from the 5TS experiment. On the other hand, they do not enable us to answer the question, posed in the previous chapter, how these results should be explained. In particular, they do not enable us to decide whether the differences between the groups is to be attributed to a difference in the extent to which the two groups make matching or half-matching responses, or tend to regard statements of the types in question as referring solely to the class of cards referred to in the subject term of the statement, or, finally, misunderstand the statement in the sense that they interpret it as implying the truth of its converse. (See page 109 above.) In the 4TS experiment, however, evidence on this point is available, at least for the two forms of the universal affirmative. (Not, unfortunately, for the F type of statement because, in truncating the F series to make it possible to complete the 4TS task in the time at my disposal I failed to make provision for this aspect.)

If a person interprets a statement of the form 'All X's are Y's' as implying that all Y's are X's, or a statement of the form 'If anything

is an X it is a Y', then he will regard as inconsistent with such a statement, not only X's which are not Y's but also Y's which are not X's. In the first set of Ah, then, as previously pointed out, he will mark as critical, not only cards 1, 4, 5 and 6 but also cards 2 and 7. Similarly with all the cards in Figure 4.2 underneath which I have placed an exclamation mark. In attempting to establish whether there existed any difference between the groups in respect of the tendency to make the illicit assumption that an A or Ah/^{statement}implies the truth of its converse I found a 'Converse Score' for each subject by counting the number of cards of the kind just referred to which that subject had marked as 'critical', awarding one half point in those rare cases where a card was marked as critical at one place in the relevant record sheet but not at the other. The results are presented in Table 4.8.

TABLE 4.8

TOTAL 'CONVERSE SCORES' OF THE TWO GROUPS ON THE AH AND A TYPES OF STATEMENT (MAXIMA = 42 AND 231 RESPECTIVELY)

	Statement Type		
	Ah	A	Both
GR group	13.0	19.5	32.5
PR group	27.0	68.0	95.0
Difference	14.0	48.5	62.5
Wilcoxon T	51	32	32
N Ranks	20	20	20
p	<.05	<.01	<.01
(two-tailed)			

It seems possible to conclude that the PR subjects were on the whole significantly more likely to make the illicit assumption that an Ah or A statement implies the truth of its converse - whether we consider these hypothetical and categorical forms of the universal affirmative separately or together. (The fact that the p value is higher in the Ah case may be attributed perhaps once again to the restriction of possible scores on this statement-type to the 0 - 2 range.) It is, I think, worth stressing the point that the interpretation to be put upon the differences between the groups in 'converse scores' seems to be certain in a sense that does not apply to other differences between them: a person who says that a Y which is not an X is inconsistent with the statement that all X's are Y's can only (accidental errors apart) be interpreting this statement as implying that all Y's are X's.

It seems, therefore, that we can say with complete confidence that the difference between the groups in the adequacy of their selections of cards in the card-turning task is at least partly to be explained in terms of the different degrees to which they are guilty of this misunderstanding. At the same time, it is not possible to say how much of the difference is to be explained in this way, and it is not, of course, possible to maintain that every PR subject is guilty of this misunderstanding or that no GR subject is. On both these last points some light is shed by the incidence of this type of mistake in each group as recorded in Table 4.8; more interesting, perhaps, are the numbers of subjects in each group who made no 'converse' errors. These are presented in Table 4.9.

TABLE 4.9

NUMBER OF SUBJECTS IN EACH GROUP WHO MADE NO 'CONVERSE' ERRORS

	Statement Type		
	Ah	A	Either
GR group	13	12	10
PR group	5	6	4

Table 4.9 confirms the superiority of the GR subject in this respect while serving to remind one, if reminding is necessary, that the misunderstanding underlying the converse error is not something which exclusively or exhaustively defines the failure of the PR subject in a deductive reasoning task.

It is, of course, to be regretted that no comparable evidence is available with respect to the F type of statement. We have already seen, from the fifth series in the 5TS experiment, that 'converse errors' were made by members of both groups in connexion with F. My own view is that they ^{are} likely to be made more often by PR subjects - mainly in view of the close logical and psychological relationship between this type of statement and the universal affirmative - but we must, of course, wait for the results of further research for definitive evidence on the point.

4.9 Results: (e) The relative difficulty of detecting 'critical cards' when the first named character is on the reverse side In the fifth series of the 5TS task one or two subjects failed to recognise the exceptions to the rule in Ah and A sets of cards. In the light of this and

other evidence relating to the Ad type of statement it was suggested that there might be some kind of 'directional set' which makes it more difficult to recognise an exception to a rule when it is in the right hand half-set of an Ah or A statement and so has the character first named on the side of the card which is originally hidden from the subject. In the present experiment the critical cards were distributed in the left and right hand half-sets as follows: Ah, 4 and 3, A, 1 and 5. Taking both groups of subjects together the total number of critical cards missed in the two half-sets respectively were: Ah, 4 and 39; A, 0 and 25. In view of the equivalence of these types of statement it seems in order to combine these frequencies - in which case the ratio of undetected critical cards in left and right hand half-sets is 4 : 64, as compared with an expected ratio of 5 : 8. This gives a χ^2 value (corrected for continuity) of 30.1 which, with 1df, has a $p < .001$.

Critical cards bearing the character first named on the side originally hidden from the subject do, then, seem to be very much less likely to be detected than critical cards with the character first named on the 'front' side. It seems possible to explain this phenomenon in terms of a directional set either at a conceptual level (of the kind we have seen Wason, 1968, invokes) or, as I have suggested, at the perceptual level. The latter offers the more direct and simple explanation since it would postulate simply a failure to recognise a critical card presented in 'reverse' order for reasons similar to those which account for failure to recognise familiar objects seen from unusual angles. On the 'conceptual' hypothesis, as I understand it, the subject's failure would be explained in terms of a preconception or misunderstanding to the effect that 'the other side' in the present situation means 'the side not at first exposed'. Other explanations are no doubt possible, of which one obvious one might be that the subject's vigilance slackens as he moves from left to right in a set of cards. This explanation, if it were the correct one, would take most of the interest from the phenomenon. It would suggest that the different degrees of certainty with which critical cards in left and right hand half-sets are detected are simply functions of the layout, and not of the 'originally-exposed-originally-hidden' dimension. This hypothesis would be more difficult to put to an empirical test than might at first appear because, of course, the fundamental factor is the order in which the cards are inspected and this would not necessarily be reversed by a mere reversal of the spatial arrangement of the cards. One final point, in connexion with the exact conditions obtaining in the 4TS experiment, is

that the equivalence of front and reverse sides of the cards ought to have been particularly hard to miss in a situation in which the characters on both sides were simultaneously in view at the end of a trial.

4.10 Results: (f) Ultimate success or failure in the verification task In the fifth series of the earlier experiment subjects were asked not only to indicate the cards they needed to turn over in order to establish the truth-value of the accompanying statement but also to turn over these cards and to say whether the statement was actually true or false. It proved impossible to say anything very definite about the relative success of the two groups in this 'verification' task because so many of them had less than all the information they needed in order to carry the task to a satisfactory conclusion - because, of course, their choice of cards had been faulty. In the 4TS experiment, however, this particular experimental defect was removed since, as we have seen, all subjects turned over all cards in the course of a trial and all had, therefore, all the information required. In this section I accordingly present data bearing on the question of success in the verification task. Table 4.10 states the frequencies with which the truth-values of the four types of statement were correctly identified. Since the maximum possible frequency in each case is equal to the number of statements of that type included in 4TS multiplied by the number of subjects involved (20 in F, 21 in the other statement-types), this is 42 for Ah and Ad, 168 for A (omitting the debatable eighth case) and 40 for F.

TABLE 4.10

FREQUENCIES WITH WHICH MEMBERS OF THE TWO GROUPS REACHED CORRECT VERDICTS ABOUT THE TRUTH-VALUES OF THE FOUR TYPES OF STATEMENT

	Statement Type			
	Ah	Ad	A	F
GR group	41	41	157	40
PR group	41	39	138	38

The differences are all in the expected direction (that is, they all favour the GR group) and the difference on A is significant on the Wilcoxon test with $T = 41.5$, N ranks = 21 and p (two-tailed) $< .01$. In fact, of course, the outcome is largely a function of differences noted in previous sections and, in particular, of the significantly greater

tendency for PR subjects to suppose, for example, that a card with a 5 on one side and a P on the other is inconsistent with a statement to the effect that all cards with a 4 on one side have a P on the other. Of the 30 erroneous conclusions drawn by PR subjects and the 11 drawn by GR subjects 23 and 8, respectively, are attributable to this confusion. Of the remainder, 3 and 2, respectively, may be supposed to be due to a failure to recognise a critical card in the right hand half-set as an exception to the rule.

4.11 Results: (g) Views about the truth-value of an A statement where the subject-class is empty. In the eighth set of A cards the statement ran as follows: 'All cards with a 5 on one side have an M on the other'. There were, however, no cards in the set with a 5 on one side. In view of the controversy amongst logicians already referred to (see above, page 117) it seemed of some interest to discover what subjects would say about the truth-value of such a statement. Two views, corresponding to the two logicians' camps seem admissible - that the statement is true (because there is no card which makes it false: the view is most plausible if the universal affirmative is to be ^{understood} in hypothetical terms, as the logicians who support it would generally maintain), or that the statement is neither true nor false (because it is neither confirmed nor refuted by the cards in the set). Subjects who take the third (and unacceptable) view that the statement is false may be confusing the truth-value of the statement with the truth-value of its apparent 'implication' (that there is at least one card with a 5 on one side)¹ or they may simply be using the term 'false' in a logically primitive way to indicate their belief that there is something wrong with the statement (meaning, perhaps, that it implies something which does not hold). Table 4.11 presents the frequencies with which the three views were taken, the 'Neither' category including a range of responses including all those in which the subject felt unable to opt for either of the other alternatives. If the 'True' and 'Neither' categories are summed, on the ground that either is defensible, a χ^2 test (with a correction for continuity) gives a value of 4.01 which, with 1df, is significant with $p < .05$. In other words it seems that PR subjects were signi-

1 If this were an implication in the full sense of the word, then of course the falsity of the implication would entail the falsity of the statement, for if p implies q and q is false, then p is false too. However, it does not appear to be correct to regard the implication in question in this way. (On different senses of 'imply' see Strawson, 1952.) Perhaps it might be maintained that PR subjects are simply less able to distinguish the different senses in which one thing may be said to imply another!

ificantly more likely to opt for the view which is not defensible. In the absence of any clear link between this difference and the relative success with which the two groups of subjects tackled a deductive reasoning task, this last result must be taken simply as one further indication of the superiority of the GR group in this general area.

TABLE 4.11

FREQUENCY WITH WHICH VARIOUS VIEWS WERE TAKEN ABOUT THE TRUTH-VALUE OF THE EIGHTH A STATEMENT

	<u>Truth-value</u>		
	True	False	Neither
GR group	10	3	8
PR group	4	10	7

Finally, in this section, it may be of some interest to report on the extent to which subjects who said the eighth A statement was true made the appropriate response to the ninth set of cards (where it was apparent, without turning over any cards, that there were no exceptions to the rule and where, therefore, subjects who said the eighth statement was true should have said that the ninth could be seen to be true without turning any cards over). In fact none of the 4 PR subjects and only 5 of the 10 GR subjects who fell into the relevant category passed what I came to regard as the ultimate test of logical perspicacity.

4.12 Review and discussion The results of the 4TS experiment have served, in some cases to consolidate, in others to extend, the conclusions reached in the previous chapter. They have shown that the responses of subjects to the card-turning task are fairly stable over a period of 12 to 18 months and (as a consequence) that differences between the groups which were significant in the earlier experiment were also significant (in every case at a higher level) in the later. One change in this respect was that the difference between the groups on the universal affirmative in hypothetical form (Ah) when both half-sets are taken together reached significance in the 4TS experiment. It was suggested, in view of the results for the other form of the universal affirmative (A) and the apparent psychological, as well as logical, equivalence of the two types of statement, that the later result was likely to come closer to representing the true state of affairs.

The main advance over the 5TS experiment consisted in showing that there was a significant difference between the groups in the frequency with which the two groups made 'converse errors', i.e., errors which could have been due only to the assumption, on the part of the subjects concerned, that a universal affirmative implies the truth of its converse. No evidence bearing on the corresponding assumption about the meaning of the F type of statement was available, though it was suggested that a similar difference between the groups was likely to exist in this respect.

In the 4TS experiment subjects were supplied with information about the correctness of their selections and exposed to a variety of combinations of cards, some of which might be expected to promote learning of the correct responses. No immediate improvement occurred in the case of the first three types of statement presented, viz., Ah, Ad and A, a fact which confirmed what Wason's researches have also demonstrated, that the errors made in the card-turning task are not easily recognised as such and corrected. In the long series of trials with the A type of statement a gradual improvement did eventually occur in the case of the PR group as a whole, and the GR group also improved rather gradually - except that, at one point, progress was unusually rapid. Large individual differences in learning were apparent within groups, perhaps the most interesting cases being those in which subjects appeared to achieve 'insight' into the source of their difficulties, realising explicitly that a universal affirmative does not imply the truth of its converse. Some fairly early improvement was apparent in the case of the remaining type of statement, F, but this was taken to be a function of its late position in the experimental series as a whole, and not of any intrinsic characteristics of the statement-type.

An aspect of performance on the card-turning task, as it relates to the universal affirmative (in either form), which was not referred to differences between the groups was the significantly greater frequency with which subjects failed to identify critical cards in the right hand half-set. The operation of a 'directional set', at perceptual or conceptual level, was mentioned as a possible cause of this phenomenon, though at least one other, less interesting, possibility had to be admitted.

The GR group's 'ultimate success' in establishing the truth-values of the statements included in the 4TS task was greater than the PR group's, and significantly greater in the case of the A type of statement, although this latter result is to be explained largely in terms of the

greater proneness of PR subjects to the 'converse error'. One final difference between the groups, the relationship of which to performance on a syllogistic reasoning task is rather obscure, was in their willingness to adopt an unacceptable view of the truth-value of an A statement when the class referred to in the subject term is empty: PR subjects were significantly more likely to say that such a statement was false.

So far as the original purpose to be served by the experimental part of this research is concerned, undoubtedly the most important single discovery has been the greater tendency of PR subjects to misunderstand the universal affirmative in the sense of assuming that a statement of this type implies the truth of its converse. The prevalence of such a mistake amongst undergraduates in general has been referred to by previous writers including, as we have seen, Wilkins (1928) and Chapman and Chapman (1959). Such novelty as may be claimed for the present research consists in the establishment of such a mistake by means of a task of a kind relatively far removed from the syllogistic reasoning in which it plays so important a role, and in establishing it as a fairly stable, and distinguishing, characteristic of the thought-processes of a group of undergraduates, selected for their superior ability and attainment, who were notably weaker on the two forms of the Valentine Reasoning Tests, Section B, than another group of undergraduates of comparable ability.

It is natural, if not exactly inevitable or even necessarily legitimate, to ask about a research finding what its practical significance is. In the present case it might be argued that even the rather informal kind of syllogistic reasoning to be found in Section B of the Valentine test plays a relatively minor role in our normal thinking and that discoveries about the source of errors in such thinking are accordingly of little importance. Whether or not the premiss of this argument is true (personally, I do not believe that it is), the conclusion does not follow, for statements of the universal affirmative type clearly do play an important role in any thinking which attempts to use or to establish generalisations - whether or not in the context of deductive reasoning. Since higher level thinking is characteristically of this kind, we might expect to find embedded in its products occasional mistakes which may be attributed to the tendency, which we have seen to exist even among subjects of the highest intelligence, to assume that a universal affirmative implies the truth of its converse. And in fact, in the space of a couple of weeks, when this possibility was in my mind, I came across

some instances which I now present.

Eysenck (1958, p. 239) suggests that the fallacy in question 'underlies all projective techniques'. Just as it would be a mistake to infer that all those who buy Jaguar cars are sporting young men from the fact (if it were a fact) that all sporting young men buy Jaguar cars, so, Eysenck says, it has been the mistake of the supporters of the projective test as a way of measuring personality to infer, for example, that all persons who make use of colour in their interpretations of Rorschach ink-blots are strongly emotional from the (alleged) fact that all strongly emotional persons make use of colour in their interpretations. Even if the premiss were true, the conclusion would not follow: 'there are many other reasons which might cause a person to be particularly conscious of the colour of the blot, and which might lead to quite different views of the subject's personality if they were taken into account.' It is unnecessary, presumably, to insist on the importance of avoiding a fallacy which, if Eysenck is right, may lie at the basis of a vast body of psychological research and practice.

My second example is one in which the author is not so explicit in his identification of the fallacy. Commenting on the failure of attempts to base items intended to test intelligence on developmental studies such as Gesell's Guilford (1967) remarks:

Perhaps it was the overemphasis upon the criterion of correlation of intelligence with age that was misleading, giving rise to the conclusion that any test that has a greater probability of being passed as age increases is therefore a measure of intelligence. Human attributes other than intelligence also increase with age; hence the correlation of a test item with age is no sure criterion of its being a measure of mental ability. (p. 11)

Clearly the error of which Guilford is speaking is of assuming that all test items more frequently passed by older than by younger children are satisfactory measures of intelligence simply because all satisfactory measures of intelligence are items which are more frequently passed by older than by younger children (there being other human attributes which increase with age). The error is precisely the 'converse error' referred to above.

My third example is also from Guilford (1967) though in this case rather more interpretation is called for on my part. According to Guilford (ibid., p. 2) Galton (than whom few are likely to be more intelligent) regarded tests of sensory capacity as satisfactory measures of

mental capacity partly because of the influence of the British empiricist school with its emphasis on the senses as the 'gateway of the mind' and partly because of what he believed to be the very poor sensory functioning of idiots. It seems to be the implication of the empiricist position that intelligence depends on the adequacy of the gateway to one's mind - so that only persons with sensory equipment of a high order can be persons of high intelligence. Similarly, the implication of the observation about idiots seems to be that their low intelligence is a product of their poor sensory functioning. In both cases the implication seems to be that good vision, hearing, etc., are a necessary condition of high intelligence ('all persons of high intelligence are persons with good vision, etc.'). Within the kind of limits suggested by the Helen Keller case, for example, this is a proposition to which we might still assent - at least if intelligence is taken to mean effective intelligence (Hebb's 'intelligence B'). On the other hand, Galton's use of tests of sensory acuity etc., could be regarded as an adequate means of measuring intelligence only if satisfactory vision, etc., were also a sufficient condition of high intelligence (so that all persons with good vision, etc., were persons of high intelligence). To this converse of the earlier proposition we should be very much less likely to assent, even with qualifications.¹

The foregoing examples seem to suggest that the fallacy of illicitly converting an A (or Ah) proposition may have been responsible for a great deal of wasted time and effort on the part of persons whose time and efforts are potentially of the greatest value. The practical importance of avoiding such errors appears to raise two further questions: first, how is it that persons of the highest intelligence should be prone to this particular error; and secondly, what steps, if any, can be taken to pre-

¹ It may be appropriate, in passing, to mention an instance from the history of psychology which relates, not to the universal affirmative but to the F type of statement, itself equivalent to the converse of an A proposition and also, as we have seen, liable to be illicitly 'converted'. The quotation is from Woodworth and Sheehan (1965, p. 116):

"Titchener had admitted or rather insisted that only well-trained introspective observers could be trusted. But Watson pointed an accusing finger at the imageless thought controversy and other recent examples of divergent results obtained in different laboratories by presumably well-trained introspectionists."

It may be, of course, that Titchener also thought that all well-trained introspectionists were trustworthy - though the acrimony with which the debate within the structuralist camp was carried on makes one doubt it. In any case, to his position as stated by Woodworth and Sheehan Watson's objection is clearly irrelevant: either Watson or the authors representing him seem to assume that 'Only X's are Y's' implies the truth of its converse.

vent or to remedy this state of affairs?

Two approaches to the first of these questions appear to be possible, at least in principle: (a) one might attempt to identify the characteristics of this type of statement, or the characteristics of human thought in general - or, of course, both - which make the mistake in question an easy one to commit; or (b) one might study subjects of high intelligence who show themselves prone to this mistake (in experimental conditions, of course) in the hope of discovering what it is about their particular capacities and habits of thought that renders them especially 'at risk' so far as this particular fallacy is concerned.

The latter of these approaches must assume, obviously enough, that there is some identifiable quality of the fallacy-prone which distinguishes them from the others; the former would presumably account for individual differences in this respect in terms of factors which are either random or, for other reasons and for all practical purposes, unidentifiable. Certainly it is difficult to know where to begin to look for the crucial factors in individual cases: there is no apparent connexion, for example, between this particular failing and emotional preoccupations of the kind that Freudians have adduced to account for other cognitive disabilities; and although (small) gaps in one's education might appear to hold out more promise as a causative factor, the fact that hardly anyone who does not take a course in logic is taught the relevant lesson formally makes it seem unlikely that the operative differences in educational experiences could ever be discovered.

As regards the characteristics of the A type of statement which might make it particularly easy to 'convert' A statements illicitly Chapman and Chapman (1959) suggest, as we saw, that their subjects had encountered deductive reasoning mainly in the context of mathematics where, they say, the converse of a true universal affirmative is generally, as a matter of fact or definition, also true. Consequently, in the abstract kind of syllogistic reasoning task the Chapmans' subjects were asked to complete, it was rather natural to assume that the converse of an A proposition was also true.

The difficulty is that we want an explanation which will account for this assumption in contexts where the thinker is unlikely to regard his activity as falling within the sphere of 'deductive reasoning' - at

least of the kind involved in mathematics. It certainly cannot be argued that the converse of any universal affirmative is usually true: when I asked a class of 41 undergraduates to write down on a slip of paper two A statements which they believed to be certainly true, they produced, in all, 61 different statements meeting this condition. Of only 5 of these could it be said with certainty that their converses were also true - one or two others being doubtful in this respect.

In an earlier chapter we saw that Woodworth (1935) attributed this fallacy to the fact that an A proposition is more 'like' its converse than it is like any other proposition. If this were the correct explanation, subjects who are prone to the fallacy would be people who fail to make the necessary discrimination between 'All S is P' and 'All P is S'. Naturally, a question then arises as to why some highly intelligent people fail to make the discrimination: the explanation does not carry us very far forward, though it does, at least, suggest a relevant piece of research, into the question whether people who fail to make this discrimination also fail to make others of a more or less similar kind.

One learns to make discriminations. If an explanation in these terms were correct, then one might hope to correct a proneness to the fallacy by suitable training. The same would not be so obviously true if the correct explanation proved to be a 'Gestaltist' alternative of the kind mentioned in my original discussion of the Woodworth position. On this view, the crucial factor might be a search for symmetry, an A proposition being asymmetrical in the sense that it states a one-way relationship between the classes of objects represented by its terms. To assume that the converse of the proposition is also true is to assume that the relationship is two-way and in that sense symmetrical (in logical terms, is a relation of equivalence rather than implication). We are again left with the question as to why the 'commitment' to symmetry should be stronger in some individuals than in others of comparable ability. It would likewise be open to the researcher to try to find some support for the hypothesis by establishing some more or less general commitment of this kind amongst the individuals concerned. The difference, as already noted, would be in the prospects of prevention or cure by training.

It would certainly be the assumption of teachers of logic and writers of books intended to help people to think more clearly (e.g., Thouless, 1945, Stebbing, 1959, and, more recently, Ruby, 1969) that it

is possible to eliminate, or at least reduce, susceptibility to the 'converse error' - and others - by means of appropriate instruction. (On this point see also Peel, 1967, p. 187.) It is true that Elton (1965) found no significant improvement in performance on the Valentine test, Section B, which could be attributed to attendance at an introductory course in logic between two attempts at the test but it is open to one to suspect that the rather diffuse lessons generally to be learnt on such a course might fail to have an effect when a concentrated effort to instil such limited points as the invalidity of 'converting' an A proposition would succeed.

The uncomfortable truth is that rather little research has been done into ways in which relatively specific logical acquisitions of this kind are made. Nor do we know, for example, what conditions of education - formal and informal - are conducive to the development of sensitivity to the relevant kind of logical mistake. Although the extensive, and highly ingenious, studies of logical reasoning by Piaget and other workers at Geneva (Piaget, 1950, Inhelder and Piaget, 1958 and 1964, for example) might be expected to contribute enormously to our understanding in this respect, there is little evidence in their writings of a concern with either inter- or intra-individual differences - or, therefore, with their causes. (On the shortcomings of the Geneva work in this respect see Hunt, 1961, p. 257 and pp. 297 f. and Flavell, 1963, p. 440.)

More that is relevant to the specific points at issue in this thesis is, perhaps, to be expected from the work of Bruner and his colleagues at Harvard (Bruner, Olver, et al., 1966). Bruner describes the development of his own interests from 'studies of individual differences in cognitive operations' (cp. his classic book with Goodnow and Austin, 1956), through a study of 'intervention and change in cognitive functioning' to his present concern with the main lines of cognitive growth. In this he admits a debt to Inhelder and Piaget and, to date, many of the papers published by this group bear a strong resemblance, in points of focus, to those which have appeared from Geneva. At the same time, there is some evidence that the Harvard researchers may owe something to the development of Bruner's own research interests, as well as to the traditional American concern with individual differences and the application of psychology to education, in their greater interest in the environment end of the organism-environment interaction which both groups regard as ^{the} fundamental continuing factor in cognitive development.

Something more directly relevant to the present inquiry is to be found in the study of children's thinking by Donaldson (1963). Donaldson worked with two groups of children of intermediate ability which she interviewed twice at an interval of two years, the younger group at ages 10 and 12, the older at ages 12 and 14. Her results underline the extent of individual differences in the sphere of deductive reasoning both in respect of an understanding of what it means to say that one thing follows from another and in respect of the further ability to tell when this condition obtains in the case of a simple syllogism or sorites. More intensive work along these lines, covering a wider range of ages and abilities, using 'syllogistic' materials presented in less verbal forms, focussing on the finer details of the reasoning process (in particular, on the understanding of the kinds of statements which play a role in such reasoning), and attempting to relate differences within and between individuals to antecedent conditions and present capacities, seems to be necessary for a fully informed answer to the question: 'Why do highly intelligent individuals commit 'converse' - and, of course, other similar - errors?'

CHAPTER FIVE

THE NEGATIVE PARTICLE AND ASPECTS OF PERSONALITY

Summary In this chapter I report an attempt to establish whether my two groups differed in the extent to which they had difficulty with the negative particle, an element of fundamental importance for reasoning in general and for the criterion Valentine test in particular, and one which Wason (1959, 1961) has shown to present difficulties for undergraduate subjects in general. Subjects were also asked to complete the Eysenck Personality Inventory, primarily with a view to explaining any difference found between the groups on the negatives task in terms of their scores on this test. Even in the absence of any such difference it was anticipated that the results from the two sources could be used to assess the validity of the underlying hypotheses about a relationship between difficulty with the negative particle and one or other of the two dimensions of personality measured by the E.P.I. The outcome of the first part of the experiment proved to be rather equivocal, there being some, rather tenuous, evidence of a difference between the groups in cases where a single negative component was involved. There was no significant difference between the groups in either of the aspects of personality measured, and only the slightest evidence of a relationship between difficulties with negatives and emotionality.

5.1 Introduction The purpose to be served by the 5TS task was, of course, the discovery of any difference which existed between the GR and PR groups in their understanding of various types of statement involved in syllogistic reasoning in general and in the criterion Valentine test in particular. One of the five types of statement was the universal negative, a type of statement, it was suggested in the light of the 5TS results, which is probably not really suited to the card-turning task since the correct response in this case is identical with the response favoured by any tendency towards 'matching'. In any case, I had decided in advance that the negative particle is so ubiquitous and important a part of reasoning (and, indeed, of language in general) that it would be useful to investigate the possibility that this part of speech, outside the context of any particular form of statement, presented greater difficulties for the PR group than for their GR counterparts. The means for such an investigation had already been provided by Wason (1959, 1961).

It is presumably unnecessary to labour the point that a difference between the groups in this respect would be highly relevant to the present investigation: it is not only that negative particles occur,

explicitly or otherwise, throughout the Valentine test, Section B, and indeed in deductive reasoning generally: two of the four types of statement recognised in the Aristotelian logic are negative in 'quality' and almost all modern systems of logic incorporate the negation sign as one of their two undefined constants, the exceptions to this rule being systems employing a single constant, such as the 'stroke' and 'dagger' functions, in which the negative is actually implicit. The notion of falsity, which, as Wason remarks, is the semantic equivalent of the (syntactical) negative particle - it is always possible to indicate the falsity of a statement by asserting its negation, in most cases simply by inserting the negative particle in the original statement - can also be seen to occupy a fundamental place in the area of deductive reasoning if it is recalled that an invalid argument is commonly defined as one in which the conclusion may be false even if the premises are true.

In his search for an explanation of the difficulties apparently presented by information set out in negative terms, Wason (1959, 1963), supported by some research by Eifermann (1961), has suggested that it may, in part, be due to the emotional response which negatives sometimes seem to elicit. In 5.4 I shall discuss this possibility at greater length and explain how I hoped to relate it to aspects of personality measured by the Eysenck Personality Inventory.

5.2 The negatives task: materials, procedure and subjects In his 1961 paper Wason describes two kinds of task in which difficulties with the negative particle appeared to reveal themselves for undergraduate subjects. In the first, the 'Verification task', subjects were presented with a series of statements, such as '2 is an even number', '3 is not an odd number', and so on, and asked to indicate the truth or falsity of a statement by pressing one or other of two buttons. There were four types of statement in all, true affirmative, true negative, false affirmative and false negative. Following the above procedure the latency of response for the different types of statement was measured with a high degree of accuracy by means of the familiar reaction-time apparatus, latencies for affirmative statements being significantly shorter than for negative ones.

In the 'Construction task' similar statements were used, except that the place where the number appeared in the verification task was left blank and the subject asked to say a number which would make the statement true or false, depending on the instruction given by the experimenter.

In this case the error component in the timing must be supposed to have been rather large in relation to the fairly short latencies involved: Wason used a stop-watch which was started as the incomplete sentence, typed on a strip of paper, was placed before the subject, and stopped as soon as the latter responded. From my own experience of using this system of measurement it seems to me that some degree of inaccuracy is inevitable, given the probable fluctuations in the experimenter's attention and the difficulty of avoiding either an anticipation of the subject's response or a time-lag after it. In general, in this task there are two reaction times involved, the subject's (which one is trying to measure) and the experimenter's (which is part of the measuring process), and it seems highly probable that the notorious variability of the former is compounded by the variability of the latter.

In view of these problems of measurement it is a rather striking fact that in Wason's research a better differentiation was obtained, between the four types of statement used, in the construction task than in the other one. In particular, it was possible for Wason to show, by means of an analysis of variance, not only that the latencies for the negative statements were longer than those for the affirmative statements but also that those for false statements were longer than those for true statements. It was because of this better differentiation between statement-types (or conditions, as I shall henceforward, more accurately, call them) that I decided to use the construction task despite the apparent difficulties of measurement.

Following Wason, four forms of incomplete statement were used: '.... is an even number', '.... is an odd number', '.... is not an even number', and '.... is not an odd number'. These were reproduced on separate strips of paper, six of each for each subject. The following instructions were read out, care being taken to ensure that they were understood:

Instructions In this part of the experiment I shall ask you to complete sentences of four different kinds so as to make them true or false. Here are the four kinds of sentence. (Here an instance of each was presented and read out to the subject.) At the beginning of each you will see a space. Your task is to complete the sentence by writing a number from 2 to 9 in the space. It doesn't matter which number you use and it will be in order to use the same numbers as often as you like.

Sometimes your task will be to complete the sentence so that it is true, sometimes you will be asked to make it false. I shall

tell you which before I lay the incomplete sentence before you.

As before (in the 5TS experiment) I shall note the time it takes you to complete the task in each case, but once again the important thing is to get the answer right. The time you take is of interest only because it indicates the relative difficulty of the different kinds of task. So please do not sacrifice accuracy to speed.

There is one minor respect in which my procedure obviously differed from Wason's: my subjects were asked to write down their responses and not simply to speak them. From an administrative point of view this change had various things to recommend it: in conducting the experiment I was freed of the need either to keep a record of responses myself or to judge the correctness of the response as it was produced, a situation in which I could concentrate on the correct operation of the stop-watch; and it was possible, before the experiment itself, to arrange the forms of statement, on their twenty-four strips of paper, in the order in which they were to be presented to any particular subject, once again making it easier to give my full attention to an essential part of the experimenter's task, the issuing of the appropriate instruction, to make the statement true or, of course, to make it false.

The order in which the four different incomplete statements were presented was combined with the two different instructions to produce an order for the four 'conditions' (true-affirmative, true-negative, etc.) which was different for each of the members within a group, the conditions appearing once in each block of four trials in a systematically varied arrangement. The arrangements for timing were as in the Wason experiment, the stop-watch being started as the strip of paper was placed before the subject and stopped as soon as he had written his response. In the case which Eifermann mentions where a subject changes his response I followed her procedure in counting only the first response - if only because the stop-watch had generally been stopped before the subject moved to make his second response.

The subjects for this experiment were the same 22 pairs as for the 5TS task, the present one having been completed at the end of a session lasting roughly an hour in which subjects had filled in Form A of the Eysenck Personality Inventory as well as taking part in the 5TS experiment.

5.3 Results (a) Errors Like Wason I found that on the whole my subjects seldom made mistakes. The incidence of errors on all six trials

of each of the four conditions is given in Table.5.1.

TABLE 5.1

INCIDENCE OF ERRORS IN THE FOUR CONDITIONS IN THE NEGATIVES TASK (ALL SIX TRIALS TOGETHER: N = 132 FOR EACH GROUP)

	<u>C o n d i t i o n s</u>				All
	True Affirm.	True Neg.	False Affirm.	False Neg.	
GR group	--	5	4	15	24
PR group	1	2	8	17	28
Both	1	7	12	32	52

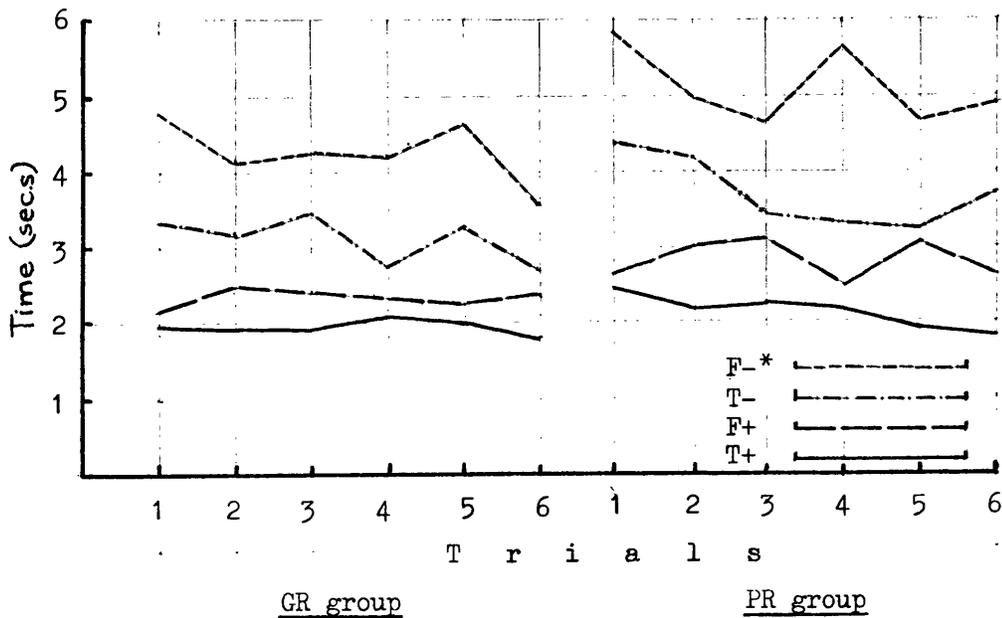
As measured by the number of errors made in each, the order of difficulty of the four conditions is the same for my PR group, and for both groups taken together, as for Wason's groups and different, therefore, as we shall see, from the order as measured by latency of response (which is the order, in terms of errors, for my GR group). As compared with Wason's groups mine made, on average over all conditions, fewer mistakes, the mean number of errors per subject per trial being .54 for Wason's subjects and .20 for my own. If the two sets of subjects are assumed to be of comparable ability in respect of this task, one might expect the drop in errors for the Glasgow groups to be accompanied by an increase in the time taken. In actual fact, a comparison of the geometrical response times of Wason's subjects and my own seems to confirm that this occurred. The longer latencies cannot, however, with certainty be attributed simply to a greater exercise of care on my subjects' parts because it must also, I think, be assumed to be due, to some extent, at least, to the fact that my subjects were required to write their responses whereas Wason's spoke theirs. In any case, the most important point to be taken from this aspect of my subjects' performance appears to be that they made few enough errors to make it possible to regard the latency of their responses as a satisfactory index of the relative difficulty of the four conditions for the different groups of subjects.

(b) Latencies of response: within groups Figure 5.1 presents the geometric mean response times of the two groups for the four conditions. I have chosen this form to facilitate comparison with Wason's results, Wason having calculated the geometrical mean, instead of the more usual arithmetic mean, because of the marked positive skew which,

as is usual with response times, at least when they are short, characterises the distributions of values in the present experiment. I shall revert to the arithmetic mean at other points in this section because the skewness of these distributions is of relatively little importance for the type of statistic to be used.

FIGURE 5.1

GEOMETRIC MEAN RESPONSE TIMES OF THE TWO GROUPS ON THE FOUR CONDITIONS



* false negative, true negative, false affirmative, true affirmative

As noted in the previous paragraph but one, the mean latencies of response for my two groups are, on the whole, rather longer than those reported by Wason - though only slightly so in the case of the GR group. In both my groups - as well as in his, with one minor exception - the mean response times for the four conditions over the six trials are consistently in the order, from shortest to longest, true affirmative, false affirmative, true negative and false negative. Needless to say, this order is confirmed when response times are averaged (arithmetically) over the six trials for each subject and comparisons made between conditions. A Friedmann two-way analysis of variance on these data, within each group, gives χ_r^2 values of 52.9 and 51.9 for GR and PR groups respectively (p in both cases being less than .001). When the differences between the conditions within each group were tested, the order given above was confirmed, the differences between successive conditions being significant even on

the low-powered Sign Test, with p in no case greater than .016. Similarly, when the times for the two false conditions were averaged and compared with those for the two true conditions, and the times for the two negative conditions with those for the affirmative conditions, the mean latency for the first member of each of these pairs was significantly longer than that for the second. Table 5.2 presents the relevant data.

TABLE 5.2

SIGN TEST DATA BEARING ON THE RELATIVE DIFFICULTY
OF THE FOUR CONDITIONS (WITHIN GROUPS)

	No. of cases (out of 22) in which the second term of the comparison had a shorter mean latency		p (Sign Test, two-tailed)	
	GR group	PR group	GR	PR
T+ versus F+	4	1	.004	<.001
T- versus F-	5	3	.016	<.001
T+ versus T-	0	1	<.001	<.001
F+ versus F-	0	1	<.001	<.001
F+ versus T-	3	5	<.001	.016
True versus False	4	1	.004	<.001
Affirm. versus Neg.	0	1	<.001	<.001

The above results confirm those reported by Wason and, in particular, his finding, as against Eifermann's, that the false affirmative is significantly easier than the true negative. (Eifermann, 1961, found no significant difference between them.) Needless to say, they also confirm the adequacy of the task, and the method of measuring the time taken to respond, as a means of differentiating between the four conditions.

Having shown, by the above means, that the order of difficulty of the four conditions was highly consistent from subject to subject within a group, it seemed natural to ask whether the order of speed of response of subjects within each group was consistent from condition to condition - in other words, whether the subjects who took the shortest time to respond to one condition also took the shortest time to respond to the other conditions. The question is one of consistency in speed of responding from condition to condition. Table 5.3 presents the relevant Spearman rhos, based on the mean response times of subjects over six trials for each of the four conditions, the rhos being corrected, where necessary, for ties (Hays, 1963).

TABLE 5.3

WITHIN-GROUP RANK CORRELATIONS BETWEEN RESPONSE TIMES
OVER SIX TRIALS ACROSS THE FOUR CONDITIONS

	<u>C o n d i t i o n s</u>					
	T+/T-	T+/F+	T+/F-	T-/F+	T-/F-	F+/F-
GR group	.64	.86	.55	.73	.61	.58
PR group	.46	.79	.32	.35	.71	.30

It will be apparent that, on the whole, the GR group was more consistent in their speed of response to the four conditions, at least when response times are averaged over six trials. The consistency of the PR group is at its greatest when the 'quality' of the condition (affirmative or negative) is held constant.

(c) Latencies of response: between groups It may have seemed odd that I have not followed Wason in using analysis of variance techniques in the evaluation of the results presented under (b). It may seem even stranger that recourse is not to made this, most powerful, statistical instrument in the present section, where an attempt is made to sort out, not only the effect of 'quality' versus truth-value, but also the effect of sorting subjects according to performance on the Valentine reasoning test: one of the more complicated analysis of variance models might appear to offer the best hope of finding an answer to the question with which the first part of this chapter is primarily concerned, whether, regardless of their truth-value, statements incorporating a negative particle present more difficulty for PR subjects than their GR counterparts.

It has to be admitted that, in modelling this part of my research on Wason's, I had assumed that a, slightly more complex, analysis of variance than the one he used would be the appropriate instrument of statistical analysis. There proved, however, to be an aspect of my results which appeared to me to rule out such a course of action. According to Hays (1963) there are three conditions which have to be met before an analysis of variance can be used with complete confidence: the distributions of scores have to be normal, the variance of different samples has to be homogeneous and the observations have to be independent. Wason was able to meet the first two conditions by means of a linear transformation of his scores. Clearly the same procedure would have been open to me and,

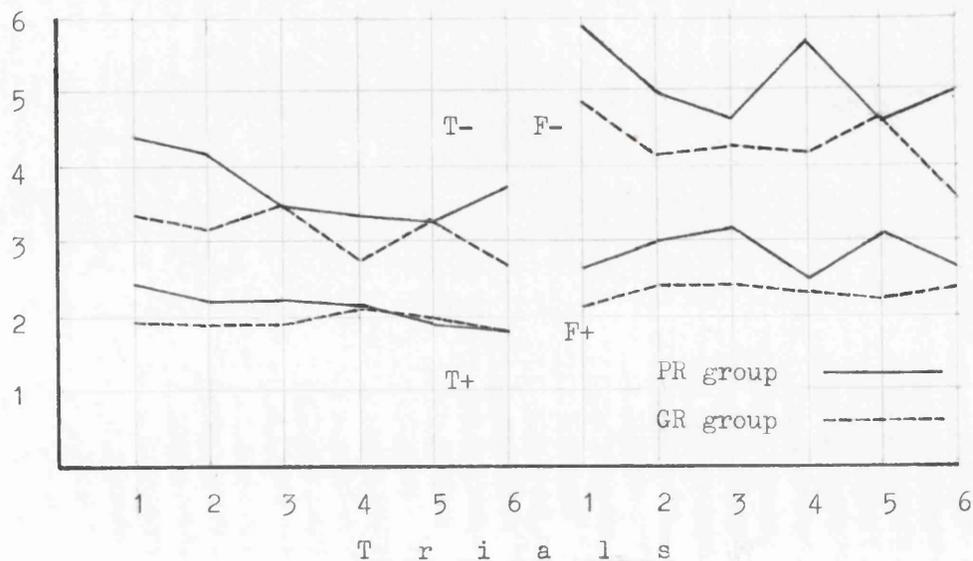
in any case, the work of Box (1953), for example, has cast some doubt on the extent to which the F test is sensitive to, at least moderate, violations of these first two conditions. On the importance of the third condition Hays lays much heavier emphasis. In particular, he expresses serious reservations about the use of analysis of variance procedures with experimental designs in which subjects act as their own controls - as, of course, they do, as between conditions, in the negatives task. (Table 5.3 underlines the extent to which the observations across conditions in the present experiment are not independent.)

It is true that Hays does not seek to exclude the use of analysis of variance in the case of any experiment in which subjects act as their own controls, but there is an aspect of my own results which make it inevitable that, in employing an analysis of variance in connexion with them, I should be violating the principle of independence of observations to a much greater extent than Wason and, therefore, I suspect, beyond the limits of the permissible. Wason's analysis was based on the reciprocals of his subjects' response times on the sixth trial, on the grounds that 'in the final phases of practice, performance would have become stable and ... errors would be at a minimum' (1961, p. 136). In other words, Wason's analysis is based on those observations which he believes to represent most validly the differences between the conditions. For convenience of comparison I have reproduced the curves of Fig. 5.1 in Fig. 5.2, juxtaposing the mean response times of the two groups on each of the four conditions. I think it will be clear from these that, whatever may be the case for comparisons between conditions within groups, comparisons between groups within conditions cannot validly be based on performances on the sixth trial.

As it happens, differences between groups are at their most favourable, to the hypothesis that the PR group experiences more difficulty with negative particles than the GR group, on the sixth trial: in both negative conditions the differences are just about as large ^{at} as/any point in the six trials, and the differences in both affirmative conditions just about as small as at any point. Unfortunately, only in the T+ case could the sixth trial be said, in any sense, to represent adequately the performances of the groups as a whole: in the fifth trials in both negative conditions the geometrical mean latencies of the two groups are virtually the same (3.26 and 3.27, 4.64 and 4.68) while, on the F+ condition, the difference between the groups is at its greatest on this trial. There is, so

FIGURE 5.2

GEOMETRICAL MEAN RESPONSE TIMES OF THE TWO GROUPS ON THE FOUR CONDITIONS
(ARRANGED SO AS TO FACILITATE BETWEEN-GROUP COMPARISONS)



far as I know, no reputable way of discounting these differences as entirely unrepresentative. In general, then, though these are, of course, the 'final phases of practice' so far as this experiment is concerned, I can not seriously claim that the performances of my subjects have become stable.

It was the apparent need to base any comparisons between the two groups on the response times for all six trials which appeared to me to rule out the use of orthodox analysis of variance techniques in connexion with my results. My subjects are 'their own controls', of course, not only across conditions but also across trials: the observations on which an analysis would be based would clearly violate the principle of independence to a very marked degree. I have depended, therefore, in what follows, on the non-parametric statistics already used extensively in this thesis.

The arithmetic mean response times for the two groups (taking all six trials together) are presented in Table 5.4, along with the relevant values for p on the application of the Wilcoxon Signed Ranks test. It will be seen that the differences between the groups fail to reach significance, on a two-tailed test, except in the case of the false affirmative condition. The result is an entirely unexpected one, not only in terms of the original hypothesis, that the PR group would experience more difficulty with the negative particle than the GR group, but also, I think,

in view of the apparently greater separation of the two groups on the false negative condition, as presented in the curves of Figure 5.2. (It is true that the F+ condition is the only one on which the PR mean response time is longer on every trial; on the other hand, one might, perhaps, have expected the rather large differences between the groups on the other five trials of the F- condition to offset the effect of the one trial on which there is virtually no difference between them.)

TABLE 5.4

ARITHMETIC MEAN RESPONSE TIMES (IN SECONDS) OF THE TWO GROUPS
ON THE FOUR CONDITIONS (N = 132 IN EACH GROUP)

	Conditions			
	T+	T-	F+	F-
GR group	2.1	3.5	2.5	5.9
PR group	2.3	4.1	3.1	6.4
Difference	0.2	0.6	0.6	0.5
Wilcoxon T	96.5	69.5	60.0	109.0
N ranks	22	22	22	22
p (two-tailed)	N.S.	N.S.	<.05	N.S.

Although it is generally unwise to comment on differences which fail to reach significance, I think it is possible to make some sense of the results presented in Table 5.4 if note is taken of the fact that the T values for the two middle conditions are of roughly the same order of magnitude while the values for the two outside conditions are very much larger. With T significant if it is equal to, or less than, 66, it will be apparent that the difference between the groups just fails to reach significance on the T- condition but does not even approach significance on T+ and F-. Now the middle conditions are alike in the sense that they both incorporate a single negative component - the (syntactic) negative particle in the one case and the (semantic) notion of falsity in the other - and it may not be entirely fanciful to suggest that it is in the presence of this condition that the two groups show a difference of response on the negatives task. This hypothesis is supported by the finding of a difference between the groups, significant at the 2 percent level on the Wilcoxon test, (two-tailed, T = 55, N ranks = 22) when an average is found for response times on the two conditions taken together. There is no other combination of conditions which produces a significant difference between the groups - and, in particular, of course, the true versus false

and the affirmative versus negative comparisons do not.

The situation is clearly being seen through a glass darkly. The darkness of the glass is partly due to the use of statistical techniques of no very great power and partly, I think, to the large error component in the measures to which these statistics are applied. The altogether tentative conclusion that there may be a difference between the groups in the ease with which they cope with conditions involving a single negative component, while it makes some sense, would obviously have made better sense if an even more significant difference had existed between them in the ease with ^{which} they coped with the condition in which there are two negative components (whereas, in fact, the T value is highest in this case). An explanation which might briefly suggest itself is that the effect of the double negative is to cancel itself out and so be tantamount to no negative at all (so that the T+ and F- conditions would be, as it were, operationally equivalent). Unfortunately, such an interpretation is inconsistent with the fact that the mean latencies for the true affirmative condition are shortest and those for the false negative condition longest. It also conflicts with Wason's finding (1961) that few subjects seemed to have adopted a 'decoding' procedure for dealing with the false negative condition, whereby a negative in the incomplete sentence was simply ignored if the preceding instruction was to make the statement false.

It would probably be pointless to speculate about the psychological processes which might account for the failure of the double-negative condition to confirm the difference between the groups apparently existing in the conditions where a single negative component is present. It is not only that this difference is itself somewhat problematic; an inspection of the response times of the two groups in the false-negative case (see Appendix H) suggests that the measure of difficulty is in this case a highly unstable one and, therefore, perhaps likely to obscure any difference which exists between the groups with respect to this condition. To obtain a crude index of the stability of the response-time measure, in this condition as compared with the false-affirmative condition, where a significant difference between the groups was obtained, I calculated the difference between the shortest and the longest response-times for each subject over the six trials. To allow for the overall difference in the magnitude of the response times in the two conditions I then noted the number of subjects in each group for whom this difference (the range, of

course) was greater than their shortest response time and also, in such cases, by how much the one exceeded the other. In the GR group there were 18 (out of 22) cases in the F- condition and 6 in the F+ condition in which the range did exceed the shortest response time of the relevant subject. Of the latter 6 cases there were only 2 in which the range was as much as twice as great as the shortest response time, while in the F- condition there were 13 such cases - with the range in eight of them being approximately 3, 4, 5, 6, 7, 9(twice), and 20 times as great. The situation is very similar for the PR group: 17 subjects in the F- condition and 9 in the F+ condition had a range which was greater than their shortest response times. Of these, 14 and 2, respectively, had ranges exceeding their shortest times by as much as twice, and of the former there were 10 cases in which the ratio was 3, 4 (thrice), 5, 6, 7, 9, 12 and 14 to 1.

In view of such wide discrepancies between the response times of individual subjects over the six trials on the false-negative condition it would, I think, be rather unwise to come to any settled conclusion about the existence or non-existence of a difference between the groups on this condition. A fortiori it would, as I have suggested, be rather pointless to try to envisage the psychological processes which might produce a significant difference between the groups in the single-negative conditions and not in the double-negative condition. Future research along these lines will have to try to achieve a more stable measure of the difficulty of these conditions or, if that should prove to be impossible, discover why the response times, in the false-negative condition in particular, are so variable. If, with a more stable measure, there still proved to be a significant difference between GR and PR groups on the single-negative conditions but not on the double-negative one, or if it proved impossible to eliminate the large fluctuations in speed of response in the F- case, it would clearly be essential to discover the psychological processes involved in responding to this task and, for this, as Wason says, it would be most helpful if a technique could be developed in which the relevant processes were 'externalised'.

5.4 The negative task, emotionality and extraversion In trying to account for the difficulties presented by a task incorporating the negative particle Wason (1959) suggested that part of the explanation might be found in the evocation of an emotional response by this part of speech. He quotes some of his subjects as saying, for example, "'Not' gave me a sort of tremor half-way through," "I don't like 'not' - it's a horrid

word," and, "The capital letters of 'not' (used in Wason's experiment) always frighten one."

The point was taken up by Eifermann (1961) who took advantage of the fact that in the Hebrew language there are two negative particles, 'lo' and 'eyno', the former of which is used in all contexts in which the English word 'not' is used while the latter is not used to express prohibitions. As Wason has pointed out (1961), the use of the English word 'not' is likely to be encountered by a child first in the context of prohibitions. The connexion between this fact and the emotional response apparently sometimes elicited by 'not' might be thought to be as follows. Part of the effect on a child of being told not to do something is likely to be an unpleasant emotional response. The scolding tone of the adult or the blow which he administers is an unconditioned stimulus producing such a response via the autonomic nervous system. Through the familiar process of generalisation, the conditioned emotional response which eventually develops may come to be made, in reduced form, whenever the negative particle, which looms so large in prohibitions, is encountered, and, in particular, when it is encountered in tasks such as the one described in this chapter. (The fact that 'not' continues to be used to prohibit certain kinds of behaviour would help to maintain the conditioned response.) It seemed possible to test the hypothesis, which I have elaborated in the foregoing, in the case of Hebrew-speaking subjects, by comparing the speed of response of subjects who completed a negatives task in which 'eyno' was the negative particle employed, with the speed of subjects for whom the statements of the task were couched in terms of 'lo'.

In the event Eifermann found that the response times of the 'lo' group were significantly longer in what Wason calls the 'verification' task - but not in the construction task. If this is taken as tentative¹ evidence in support of the hypothesis that difficulties with the English word 'not' are partly to be attributed to the unpleasant emotional response which it elicits, then, as Eifermann suggests, we should expect greater difficulties (longer response times) in the case of subjects whose emotional responses are greater. She suggests an experiment in which a group of

¹ Tentative not only because of the inconsistency of her results from the two kinds of task but also because, as Eifermann says (p. 268), the connotative difference between 'lo' and 'eyno' is different also in the sense that 'the information contained in a 'lo' sentence may be more easily categorized as belonging to a particular context, since 'eyno' does not appear in prohibitive contexts'.

'emotional' subjects would be compared with a group of 'non-emotional' subjects in respect of their response times in a negatives task. I think the opposite approach would also be possible: the emotionality of subjects with longer response times on a negatives task could be measured and compared with the emotionality of subjects with shorter times. It was partly because of the expectation that my PR group would prove to have a significantly longer mean response time on the negative conditions of the construction task that I decided, in advance, to measure their emotionality by means of the N scale of the Eysenck Personality Inventory.

At the same time, it seemed to me arguable that difficulties with negatives, as measured by speed of response, would be related, not, or not only, to the emotionality (or 'neuroticism') of subjects but to their degree of introversion. The reasons for this view are to be found in Eysenck's account of the socialisation process (1964, for example), coupled with his contention that introverts condition more easily and permanently than extraverts (ibid.). The relevant part of Eysenck's view of the process of socialising a child has already been sketched, in the last paragraph but one, where, it was suggested, the parent attempts, consciously or unconsciously, to eliminate undesirable kinds of behaviour by associating it with consequences which are unpleasant - in the first place, usually, punishment or scolding but, in the long run, at least, simply with the unpleasant emotional reaction in the child which accompanies the harsh words or deeds. The paradigm, Eysenck suggests, is classical conditioning and its effectiveness is to be seen, not only in the adequately socialised individual with his well-developed conscience (the continuing activity, or activation, of his autonomic nervous system) but also in the neurotic in whom an emotional response has become attached, by accident or by generalisation, to an inappropriate situation, object or action.

Differences in the extent to which people are adequately socialised are to be explained, on this theory, sometimes in terms of the thoroughness with which the original conditioning process has been carried out, sometimes in terms of the person's position on the extravert-introvert continuum. Because, in general, introverts, with their low levels of inhibition, condition more readily than extraverts, given the same amount of socialisation the former are likely to be better socialised than the latter: their emotional responses to prohibited behaviours, to prohibitions themselves or to the words and gestures used in prohibitions are likely to be stronger. To the extent that difficulties with negatives are due to

the role they play in prohibitions, then, one might expect a group of introverted subjects to have more difficulty (make slower responses) in a negatives task than a group of extraverts. Similarly, if a group of subjects with a significantly greater mean response time in a negatives task proved to be also significantly more introverted than another group of subjects, one might infer that the latter fact provided at least a partial explanation of the former.

The Eysenck Personality Inventory, with its scale for extraversion as well as for emotionality or 'neuroticism', appeared to be eminently suitable as an easily administered and brief measure of these aspects of the personalities of my subjects. If they turned out to be significantly different in their speed of response to the negative statements in the construction task, at least a partial explanation might be forthcoming in their different degrees either of extraversion or emotionality. If there was no significant difference between them on the negatives task but one on the E or N scales of the E.P.I., this would obviously still be an interesting result, although with no obvious explanation (related to no prior hypothesis).¹ Finally, in the absence of significant differences between my groups in either respect, it would still be possible to look for a relationship between degree of extraversion or emotionality, on the one hand, and speed of response to the negative condition on the other: evidence bearing on Eifermann's hypothesis and my own would still be available, though, of course, not via my two groups as originally constituted.

5.5 The Eysenck Personality Inventory : general aspects The Eysenck Personality Inventory is too well known for it to be necessary for me to say very much about it. An improved version of the Maudsley Personality Inventory, it incorporates a Lie Scale, for the detection of 'faking', as well as the E (extraversion) and N (neuroticism) scales already referred to. The choice of these two dimensions for inclusion in the inventory reflects the Eysencks' belief (Eysenck and Eysenck, 1964) that these two factors 'contribute more to a description of personality than any other set of two factors outside the cognitive field', a belief based on extensive factor analytic studies in the field of personality description.

The Inventory itself is available in two Forms, A and B, in each of which there are 57 questions, 24 for each of the E and N scales and 9 for the Lie scale. Subjects are required to choose between 'Yes' and 'No'

1 Evans (1964) found a significant negative correlation between N and performance on the Valentine test as a whole (i.e., both Sections).

in answering the items, the intermediate 'Don't know' or '?', sometimes included in such inventories, having been omitted in order to combat the 'response set' of answering a large number of questions in this non-committal fashion. One curious feature of the Inventory, in my view, is that all the items in the N scale are keyed in the same direction - so that a 'Yes' answer to such an item always scores a point towards the N scale total. Since the tendency to answer all or most of the questions in an inventory in the affirmative - to 'acquiesce' - is a well-established characteristic, distinguishable from neuroticism (Vernon, 1965), it is strange that this aspect of the E.P.I. has remained unchanged. The Eysencks say, in the Manual, p. 13, that the acquiescence response set has been investigated 'rather intensely' in relation to the M.P.I. and the E.P.I. and has been found to play a 'very small and unimportant role only'. With the M.P.I. study (Eysenck, 1962) Vernon, at least, is not satisfied and the Eysencks do go on to admit that acquiescence 'is not completely absent, however, and may require to be borne in mind with certain groups of subjects'. In the absence of further guidance in the matter, it may be necessary to bear it in mind in connexion with my own groups of subjects.

Test-retest reliabilities for two rather small groups of subjects, over periods of a year (N=92) and nine months (N=27) are given in the Manual. They range from .80 for the E scale, Form B, to .94 for the same scale, both Forms together. Intercorrelations between the two Forms, completed in a single session, and based on the answers of 2000 normal subjects, are given as .76 and .81 for the E and N scales respectively. Validity data in the manual is somewhat general in character, the reader being referred to other publications for evidence that the results of the tests 'fit in with predictions made from a more general theory' and that 'individuals who impress others as showing introverted or extraverted behaviour patterns, or as being stable or unstable in their everyday behaviour, answer the E.P.I. in a corresponding manner'. Norms are provided for 22 different categories of person, one of which is students (N = 347 of whom 158 were male), in the form of means and standard deviations. Eysenck believes that one Form of the Inventory may be sufficient 'for experimental studies', though the two Forms make it possible to retest after experimental treatment 'without interference from memory factors'. These last two points have a, somewhat indirect, bearing on some points I shall be making in connexion with my own use of the Inventory.

5.6 Eysenck Personality Inventory: administration and results

Subjects are instructed to work quickly through the Inventory, giving as their answer to a question their first thoughts on the subject. Apart from its other advantages, this instruction means that one Form of the Inventory can be filled in in a very short time. My subjects completed Form A during the first experimental session, after the 5TS task and before the negatives task, a subsidiary purpose to be served by this arrangement being to give subjects a brief rest between the two more taxing tasks. They completed Form B 12 to 18 months later, in their second experimental session, the subsidiary purpose to be served on this occasion being, as already mentioned in connexion with the 4TS task, the reduction of transfer effects from the A type of statement (which preceded it) to the F type (which followed).

It is not altogether/^{clear} how much weight can be placed on my subjects' responses to Form B in view of the fact that Form A was scored and its results explained at the end of the experimental session. Thus subjects might be expected to remember, on the second occasion, the kind of aspect being measured and also, of course, the existence of the L scale. A priori one could not be certain how much effect this knowledge was likely to have: subjects had, of course, no detailed knowledge of the Inventory or, therefore, of the items which would belong to the different scales, and they had relatively little reason to try to improve upon their apparent personalities, even if they had known how to achieve this, since the purpose of the testing was carefully explained and, I think, generally accepted.

Looking at the results from the two Forms themselves for evidence of an effect gives no unequivocal answer to the question of the comparability of the results. On the one hand, as Table 5.5 shows, there was a noticeable drop in L scale scores on Form B and the product-moment correlations between the two Forms are only .51 and .31 for the E and N scales respectively in the case of the GR group. On the other hand, the corresponding values for the PR group are .70, a result which seems altogether satisfactory in view of the fact, noted in the previous section, that the intercorrelations between the two scales when the two Forms were completed, by the standardisation samples, in a single session, are reported in the Manual as .76 and .81 respectively. Finally, an increase in E scale scores for both groups which might naturally be interpreted as a sign that my subjects had succeeded in shifting their scores in the direct-

ion of greater extraversion actually reflects fairly accurately the difference in the means of the student standardisation sample: the standardisation sample's mean increased by 2.34 from A to B, the means of my GR and PR groups by 2.47 and 1.67 respectively. The same applies in the case of the N scale means (where the standardisation sample's means were 10.00 on A and 11.04 on B). In the case of the N scale, however, there is a discrepancy between the change one might expect from the norms and the change which actually occurred: whereas the standardisation sample's standard deviation fell from 5.01 to 4.82 from A to B, the corresponding values for both my groups show an increase. If this discrepancy calls for an explanation, however, I think it is more likely to be found in the immediate context in which my subjects completed Form B than in any effect of the scoring and discussion of Form A. The later Form, as already mentioned, was completed at the end of the long series of trials on the A type of statement in the 4TS task, in the course of which some subjects had met with notable success and others with equally notable failure. It is perhaps not entirely fanciful to suppose that the differences in emotional response visibly evoked by such different degrees of success might be reflected in the answers to a scale designed to test emotional stability.

TABLE 5.5

MEANS AND STANDARD DEVIATIONS OF THE TWO GROUPS ON THE EYSENCK PERSONALITY INVENTORY, FORMS A AND B (N = 21* FOR BOTH GROUPS)

		Form A			Form B			Forms A + B		
		E	N	L	E	N	L	E	N	L
GR group	Mean	11.09	11.90	1.86	12.76	12.71	0.19	23.85	24.61	2.05
	S.D.	3.49	3.22	1.42	3.11	4.57	--	5.74	6.33	--
PR group	Mean	9.67	13.38	2.48	12.14	14.86	0.38	21.81	28.24	2.86
	S.D.	3.88	3.06	1.47	2.66	3.98	--	6.09	6.09	--
Mean diff.		+1.42	-1.48	-0.62	+0.62	-2.15	-0.19	+2.04	-3.63	-0.81

* I have not included the scores of the pair of subjects, the PR member of which was not able to attend for the second experimental session.

The question of the admissibility of the Form B results is, in any case, largely an academic one, of more interest, perhaps, to the student of the E.P.I. than relevant to the questions at issue in this part of the

present thesis. The truth is that, while the differences between the groups on the E and N scales are in the expected direction in both Forms, none of them is large enough to be significant even on the parametric t test, which it seemed appropriate to use with scores on an Inventory presumably constructed so as to give a normal distribution of scores. This means that any difference which may be thought to exist between the groups in the amount of difficulty experienced with negatives cannot be explained in terms of a difference between them in degree of extraversion or emotionality, at least as measured by the Eysenck Personality Inventory. In so far as differences in strength of emotional reactions are tied to such differences in personality these too must be rejected as sources of differences between the groups with respect to negatives. And finally, of course, any other, less specific, relationship between logical reasoning ability, as measured by Section B of the Valentine tests, on the one hand, and extraversion or emotionality on the other must also be discounted.

The last point to be looked at in connexion with the E.P.I. results concerns any direct evidence of a relationship between difficulties with negatives and scores on the E and N scales, it being possible that such a relationship exists even if it does not account for the differences between the GR and PR groups. To investigate the possibility I calculated rank correlations, corrected for ties, between the E and N scores of all subjects (both Forms together) and mean latencies over the six trials on the four conditions of the negatives task. These are presented in Table 5.6.

TABLE 5.6

RANK CORRELATIONS BETWEEN E AND N SCORES OF ALL SUBJECTS (FORMS A AND B)
AND MEAN RESPONSE TIMES ON THE FOUR CONDITIONS OF THE NEGATIVES TASK

	<u>Conditions</u>			
	T+	T-	F+	F-
Extraversion	.03	.10	.01	.08
Neuroticism	.18	.32	.27	.00

There is, of course, a difficulty about establishing the significance of a rho coefficient - especially in the case where some ranks are tied (Hays, 1963, p. 646). In view of the sample size, however, it may seem legitimate to adopt the t test procedure described by Hays. In that

case the only correlation in the above Table which is large enough to reach significance is that between the latency of response to the T- condition and neuroticism. If this can be taken to mean that there is a (not very pronounced) tendency for more emotional subjects to take longer to respond to this negative condition than less emotional subjects, it may be regarded as tentative evidence in support of Eifermann's hypothesis as described in section 5.4. The fact that the relationship is not a stronger one may be attributed to the operation of other factors in determining speed of response in this task - and, of course, both Eifermann and Wason have insisted that the hypothesised emotional component in responses to the negative particle is only one possible source of the difficulties people have with this part of speech. The evidence for a relationship between emotionality and speed of response to the negative conditions of the construction task would have been stronger, of course, if there had also been a significant correlation between scores on the N scale and the false-negative condition. On the other hand, I have already remarked on the instability of the response-time measure in this case.

Two things remain to be said. The first is that there has been no evidence at all to support my attempt to relate latency of response in the negatives task, via the different conditionability of subjects, to differences in extraversion. The other is that a more direct measure of emotionality, coupled with a better measure of difficulty with negatives, might be expected to produce a clearer test of the role of emotional responses in a subject's attempts to deal with information couched in negative terms.

5.7 Summary In the present chapter I have described attempts to establish whether there was a difference between the GR and PR groups in respect of difficulties experienced with the negative particle (a fundamental factor in syllogistic reasoning) and in degree of extraversion or emotionality, as measured by the Eysenck Personality Inventory.

Though the Wason 'construction task' employed for the first of these purposes proved adequate to discriminate between the four conditions within groups and to confirm the findings of Wason in this respect, the result of the comparison between groups, within conditions, was something of a surprise and, it was suggested, might be partly due to the instability of the response-time measure of difficulty, especially in the false-negative condition. If this were the case, then the complete absence of any

significant difference between the groups in respect of the condition just mentioned might not be thought to represent a fatal obstacle to the conclusion that there was a significant difference between the groups, not specifically in respect of the speed with which they responded to statements involving the negative particle but, more generally, in conditions where they were required to cope either with the negative particle or with its counterpart at the semantic level, the notion of falsity. At best, however, such a conclusion must continue to be regarded as somewhat problematic: although the differences between the groups on the false affirmative condition by itself, and on the false affirmative and true negative conditions taken together, were significant (respectively at the 5 and 2 per cent levels on a two-tailed test), the difference on the negative condition by itself just failed to reach significance.

Differences between the groups in degree of extraversion and emotionality were looked for primarily as an explanation of any significant difference which might be found between them in respect of the negative particle, the connexion between the two being supposed to be via the emotional response to this part of speech reported by some of Wason's subjects and tentatively and indirectly confirmed by Eifermann. It was considered that such a response might be related either, as Eifermann had suggested, to degree of emotionality or, as the present writer argued, to degree of extraversion via the early exposure of individuals to the use of the negative particle as a prominent part of prohibitions. In the event there were significant differences between the groups in neither of these respects, as measured by the Eysenck Personality Inventory, so that any difference which existed between the groups in respect of the negative particle (or of the negative components of a task) could not be attributed to this source. At the same time, a small, but significant, relationship was shown to exist between speed of response to the true negative condition of the construction task and scores on the N scale of the E.P.I., providing some tentative evidence in support of the Eifermann hypothesis.

CHAPTER SIX

REVIEW, DISCUSSION AND PROSPECT

The purpose of the research described in this thesis was to discover some of the factors responsible for failure in a syllogistic reasoning task such as the one represented by the Valentine Reasoning Tests, Section B. To this end a review was made of the literature bearing on the sources of error in reasoning of this kind, there being virtually nothing specifically on the question of individual differences. The most widely accepted theories about the factors producing errors in syllogistic reasoning were found to be (1) that the 'atmosphere' of the premises influences the subject's choice of conclusion (in the traditional multiple-choice type of syllogism test) and (2) that syllogisms couched in terms of 'emotionally significant' material are less likely to be evaluated correctly than syllogisms whose content is 'emotionally neutral'. Other possibilities suggested in the literature were (3) that syllogisms with conclusions which are judged to be true are likely to be thought to be valid and syllogisms with conclusions judged to be false invalid and (4) that errors might be due to a failure to understand the nature of the task or to abide by its terms if these were understood in the first place. It had also been noted (5) that there is a tendency for subjects to misunderstand some of the types of statement involved in syllogistic reasoning, in particular the universal affirmative (which is understood to imply the truth of its converse) and the particular affirmative and negative (the 'some' being interpreted as meaning 'not all').

Although these five sources of error were supposed to be characteristic of the (usually) undergraduate population at large, each could obviously be made to generate a hypothesis about the source of individual differences on the general assumption that some subjects are more prone to the relevant effect than others. However, the reality of the effect had, of course, to have been established and it was argued that in the case of the first two this was not so. The 'atmosphere effect' does draw attention to, and offer an explanation of, a pattern in the choice of erroneous conclusions but it cannot be thought to explain (and I think it was not originally intended to explain) why erroneous conclusions are preferred to correct ones. On the contrary, properly understood, it is

all to the good to be influenced by 'atmosphere' for the 'secondary hypotheses' in terms of which it is applied to syllogistic arguments by Sells are two of the 'rules of the syllogism', as he himself recognises. To be guided by these in one's choice of conclusions is clearly better than to be guided by none - although not, of course, as good as being guided by all, including, in particular, those relating to the 'distribution' of the terms.

Papers purporting to establish emotionally significant content as a factor reported researches in which at least one other important variable had not been controlled. Of these variables one was the relationship between the truth-value of the conclusion and the validity or invalidity of the argument, the factor mentioned under (3) above. A study such as Thistlethwaite's which compares the performance of a prejudiced group on a set of arguments touching on their prejudices with the performance of an unprejudiced group on the same material is liable to founder on the difficulty of distinguishing the effects of any emotional components of prejudice from the effect of the different views about the truth-values of the conclusions of the arguments in which the difference between being prejudiced and unprejudiced at least partly expresses itself - it having been established by Wilkins and by Janis and Frick that an argument tends to be judged valid if its conclusion is true and invalid if its conclusion is false even in circumstances where emotion cannot seriously be thought to play a part.

There was evidence, from an inspection of the responses of the poor and good reasoners to the two Forms of Section B of the Valentine test, that they might differ in respect of their susceptibility to the last three of the five effects listed above. There were large differences between the groups on those items where the argument was valid and the conclusion judged by another, comparable, group of undergraduates to be false. In other arguments where the two groups differed substantially in their degree of success, PR subjects showed a greater tendency to offer as a reason for rejecting an argument the falsity of one of the premises (showing that they did not understand or - more likely, in view of their success with other items - did not consistently adhere to the terms of the task), or appeared to suppose that 'Only X's are Y's' implies the truth of its converse, the mirror-image of the confusion about the meaning of the universal affirmative referred to above.

It was in this last area that the efforts of the experimental part of the research reported in this thesis were concentrated. Having selected two groups of subjects, matched one to one on a composite measure of academic attainment and ability but differing widely in respect of their scores on the two Forms of Section B of the Valentine Reasoning Tests, I tried to establish first of all whether they differed in the adequacy with which they grasped the meaning of five types of statement, three of them playing prominent roles in the items of the criterion tests and two playing smaller parts but incorporating the logically important connectives 'if ... then' and 'either ... or'. These were the universal affirmative in categorical and hypothetical form (A and Ah respectively), the universal negative (E), a universal-disjunctive (Ad) and the 'Only X's are Y's' form of statement (F).

The assumption on which the 'five types of statement' (5TS) experiment was based was that differences in performance on Wason's 'card-turning' task, as applied to the five types of statement in question, would indicate differences in the extent to which GR and PR subjects understand these types of statement. Since PR subjects did significantly less well on the A and F types but not on Ad, E or (with one exception) Ah, the inference would appear to be that PR subjects understand the A and F types (but not the others) less well than their GR counterparts. Such an inference seems to be supported by the results of the later, 4TS, experiment where the differences between the groups, in slightly different circumstances, proved to be highly significant, not only in A and F but also, and as one should expect, on the other form of the universal disjunctive Ah, and where the groups differed in their proneness to the 'converse' error, an error which admits of only one interpretation, namely, that the subject who commits it misunderstands the statement concerned (in the 4TS experiment the universal affirmative) in the sense that he assumes that the relationship said to exist between its terms is a reciprocal, and not simply a one-way, one. On the reasonable assumption that a similar difference between the groups could be established in their proneness to commit the converse error in the case of F statements, it seems as if one may conclude that a difference has been shown to exist between the groups which is of considerable explanatory value in respect of their performances on the Valentine test, and, by implication, of their deductive reasoning ability in general - for we have noted that the illicit 'conversion' of the universal affirmative would render a subject liable to some of the fallacies which Wilkins found undergraduates had most diffic-

ulty in detecting and which were most effective in discriminating between those who scored high and those who scored low in her syllogism test. Equally, at a more specific level, we have seen that this mistake, and its counterpart in the case of F, would account for some failures in the criterion Valentine test.

It is, however, necessary to admit that the above interpretation of the differences between the groups on the 5TS (and the corresponding part of the 4TS) task, though in my view the most plausible, is not an inescapable one. This is true, I think, even of the connexion between the significant differences in the number of converse errors made and in success in recognising the cards which need to be turned over to establish the truth or falsity of an A statement: although, as I say, the latter failure may very plausibly be attributed to the former, it is open to a critic to suggest that the difference between the groups on this statement type in the 5TS task would not have reached significant proportions in the absence of other predisposing factors. What some of these other factors may be became apparent in the discussion of various aspects of the responses to the 5TS task and it is the probable intervention of these factors which throws the strongest doubt on the alleged relationship between success on the 5TS task (and its 4TS counterpart) and adequacy of understanding of the statements involved.

Perhaps the most important single piece of evidence in this connexion was drawn from the 1968 paper by Wason and Johnson-Laird where it was reported that subjects make fewer errors on the card-turning task when they are asked to deal with the universal affirmative in its disjunctive form than when dealing with the hypothetical (and, by implication of Wason's own results as well as those reported in this thesis, the categorical) equivalents. To insist that the first of these is not easier to understand than the other two is admittedly to set a subjective estimate against what may appear to be an objective measure of difficulty, but this is a procedure which has characterised the development of objective measures of subjective phenomena since the beginning of psychology and it may be supported in this particular case by reference to Wason's own demonstrations of the difficulties presented by information in negative form - for the disjunctive equivalent of the universal affirmative is 'Everything is either not X or it is Y' whereas the categorical and hypothetical forms, with which we have become very familiar in the course of this thesis, are, 'All X's are Y's' (or, 'Every X is a Y') and 'If any-

thing is an X it is also a Y'.

As I have said, the implication of this case, that success on the card-turning task does not depend in any straightforward way on the intrinsic difficulty of the statement in question, is supported by the the results of the progressive analysis of errors in the 5TS task. Consideration of the location (left or right hand half-set) and nature (commission or omission) of the errors most frequently occurring in different statement-types suggested that the success with which subjects tackle the card-turning task with reference to any particular statement-type depends on the extent to which (a) the correct response approximates to a 'matching response' (in which the subject turns only the cards with named characters), (b) the correct response is the same in both half-sets, and (c) the grammatical form of the statement encourages the belief that it refers to both classes of cards mentioned and not simply to the first.

The operation of these factors, in addition to the one suggested by Wason's and Johnson-Laird's results referred to above - that success on a statement-type depends on the extent to which the grammatical form of the statement alerts subjects to the difficulties of the task - makes it certain that the card-turning task cannot be used as a means of assessing the relative intrinsic difficulty of grammatically different types of statement. It does not, however, follow that differences in the success with which individuals respond to the card-turning task as it relates to any particular type of statement are not largely determined by differences in their grasp of the meaning of statements of that type: otherwise, of course, the view of the results of the 5TS experiment which I have described as highly plausible would not be tenable. At the same time it clearly requires one to assume that there are no significant differences between individuals in respect of the four factors I have mentioned as affecting performance on the card-turning task or else that such differences as exist are closely associated with differences in understanding of the statement-type in question. In fact I have tried to show that the second of these alternatives applies in the case of the significant difference which was found in the frequency with which GR and PR groups made matching responses in statement-types where these are not identical with the correct response (as they are in E): a subject who supposes that 'All X's are Y's ' means or implies that all Y's are X's or that 'Only X's are Y's' means or implies that only Y's are X's would naturally express this view in a tendency to make matching responses.

If these assumptions are not granted, however, the significant differences between the GR and PR groups on A and F statements in the 5TS task, and on these and Ah in the corresponding part of the 4TS task, would most naturally be taken to indicate differences bearing no explanatory relationship to success or failure in the syllogistic reasoning task, and the inescapable fact, established in the 'converse errors' component of the 4TS task, that PR subjects are more likely to assume that a universal affirmative implies the truth of its converse, would represent a rather fortuitous confirmation of a hypothesis about the sources of differences in syllogistic reasoning ability. On this view the differences between the groups in the 5TS experiment and its 4TS counterpart would simply be another way in which the differences between the two groups in their ability to cope with deductive reasoning tasks set out in predominantly verbal terms expresses itself and the fact that the differences occur in the cases they do and not in Ad and E, far from reflecting a highly significant relationship between the statement-types involved, would simply be a function of the greater difficulties these statements present.

As I have said, this is not the view I myself take: I present it because it suggests itself as a plausible alternative to my own view and because I have not as yet collected the kind of evidence which would be necessary to decide finally between the two. Whether or not we interpret the differences between the groups on the 5TS task as reflecting differences which could partly explain their differing degrees of success on the Valentine test, Section B, or ^{simply} as further evidence of the GR group's superior deductive reasoning ability, there have been aspects of the responses of the two categories of subject which have seemed to merit some attention even though their bearing on performance in the Valentine test seemed obscure from the first. These include the evidence of 'playing safe' on the part of the GR group in the 5TS task, the speedier progress which this ^{group}/showed in the nine trials with the A statement-type in the 4TS task (with its incidental confirmation of the importance, for success in this task, of not assuming that a universal affirmative implies the truth of its converse) and the significantly greater frequency with which the PR group took an unacceptable view of the truth-value of an A statement which refers to an empty class. Into this category too, though at the furthest remove from the central problem of this thesis, because not apparently related to the GR-PR dimension, comes the observation of the greater difficulty subjects had in recognising a card as an exception to a rule if it had the named character on its reverse side.

What has been said about the difficulty of being certain about the correct interpretation of responses to the card-turning task serves to illustrate the point that an experiment in psychology - and particularly, of course, one employing fairly novel procedures - is likely to reveal unsuspected aspects of the task as well as reflecting the capacities of subjects: progress in assessing the latter is made by taking account of the former, as was done, to some extent, in the evolution of the 4TS experiment from the earlier 5TS one. The same duality of outcome may be seen in the case of the 'negatives task' with which the first part of chapter 5 was concerned. Originally intended to establish whether PR subjects have more difficulty than GR subjects with negative particles the results seemed to suggest a difference between them with respect to a single negative component, whether the component were 'syntactical' (as in the need to deal with a statement couched in negative terms) or 'semantic' (as in the condition where subjects were required to make a statement false). At the same time, the instability of response times in the condition where both these components were involved made it impossible to be certain whether the confirmation naturally looked for in this quarter failed to materialise because the difference between the groups in respect of negative components is not a real one or because the measuring instrument is defective: an aspect of the experimental situation which could go unnoticed, or unremarked upon, as long as the experimenter's purpose was to make a comparison within groups across conditions becomes obtrusive when the aim is to compare performance within conditions across groups. As I have suggested in chapter 5, an attempt to achieve a more stable measure of the difficulty of the false-negative condition is important/^{not only}for the kind of purposes served by the research reported in this thesis but also for the investigation of the Eifermann hypothesis of a relationship between emotionality and difficulties with negatives: as it is, the discovery of a small but significant relationship between scores on the N scale of the E.P.I. and slowness of response to the true-negative condition must be regarded as a somewhat problematic source of support for that hypothesis in the absence of any relationship between N and/^{speed of}response to the false-negative condition.

The extraversion and neuroticism or emotionality of the GR and PR groups was assessed by means of the Eysenck Personality Inventory, primarily with a view to testing the hypothesis that a difference between the groups in the difficulty experienced with negatives might be explained in terms of a difference in one or other of these dimensions of personality.

In fact, although PR subjects were on average more introvert and more emotional ('neurotic') than their GR counterparts, as the relevant hypotheses predicted, neither difference was large enough to be significant on a two-tailed t test. Hence it seems likely that the difference in deductive reasoning ability on which this thesis has focussed is not related to a difference in these aspects of personality either directly, or indirectly, via the ubiquitous negative components of deductive reasoning tasks or via some other, unspecified, factor or factors.

An aspect of the outcomes of the various comparisons of the two groups made in the course of this thesis which should not go without comment in this concluding review is the extent of overlap between the groups, a striking concrete instance of which would be the performance, in the 5TS task, of the PR subject, LD, which excelled all others including the best of the GR subjects' performances. Of course it would be naive to suppose that the 'GR' and 'PR' labels correspond to any clearcut division of subjects into classes. It is not only that the instruments used to constitute the two groups are of imperfect reliability and validity but also that the basis of selection was such that while the average standing of the two kinds of subject on the criterion Valentine test was markedly different, the range of scores within each group was such that the lowest GR score exceeded the highest PR score by only one T-scale point. What was fairly constant from one pair of subjects to another was the gap between performance on the Valentine test, Section B, and the academic attainment and ability of the PR subject as measured in other ways. It ought not to come as a surprise, therefore, that some PR subjects never made a converse error in the 4TS experiment and that some GR subjects did make mistakes of this kind. The conclusions drawn from any experiment in the biological or social sciences is likely to be statistical in character: it is just that certain events or characteristics occur so much more often in one set of conditions or in one kind of organism than they do in another that it is highly unlikely that there is only an accidental relationship between the difference in frequency of occurrence and the difference in condition or organism. Accordingly, the main conclusion supported by the experimental part of this thesis appears to be that when general academic attainment and ability is held relatively constant, different degrees of success on a deductive reasoning task are associated in subjects of high ability with different degrees of proneness to assume that the converse of a true universal affirmative is also true - and (probably) that the converse of a true F proposition is also true - and,

again with some probability, with different degrees of difficulty with the negative components of the kind of task in question.

The research reported in this thesis, like most researches, is incomplete in the sense that it leaves many questions unanswered, including some which have been raised by the findings of the present study and one or two upon the answers to which rests a decision about the proper interpretation of the findings themselves. Into the last category, as we have seen, fall (a) the question whether the absence of any significant difference between GR and PR groups on the false-negative condition of the negatives task is attributable to the instability of the response-time measure of difficulty in this case or the absence of any real difference between the groups on this condition and (b) the question whether the GR and PR groups differed significantly in the success with which they responded to the A, Ah and F types of statement (but not to Ad and E) because of the meanings of the statements concerned and their roles in a syllogistic reasoning task or whether the difference is to be attributed simply to a difference between the groups in reasoning ability in general, together with the greater difficulties presented, specifically in connexion with the card-turning task, by the A, Ah and F types of statement. Answers to both questions depend, in turn, on the development of appropriate measuring instruments, obviously, in the former case, a stable means of establishing the difficulty different individuals experience with the false-negative condition, less obviously, perhaps, in the latter case, a way of determining the amount of difficulty a subject has in understanding a particular type of statement (the non-existence of any obvious means of achieving this having prompted the use of the card-turning task in the first place). One other task, again of the most immediate relevance to the conclusions drawn in this thesis although unlikely to present practical difficulties of the same dimension, is the investigation of the incidence of 'converse errors' in the F case and their relationship to success and failure in a syllogistic reasoning task.

Other matters, arising from the research reported in this thesis but not of such immediate relevance to the problem which it investigates, concern the existence, and mode of operation of, a 'directional set' which hinders the recognition of exceptions to a rule in the card-turning task when the cards in question have the character first named on their 'reverse' sides, and the incidence and basis of the view that a universal affirmative which refers to an empty class is false.

The second of these questions touches on what I have described, at the end of chapter 4, as our ignorance about some of the fundamental components of the logical appraisal of arguments, in particular, in this case, the extent to which undergraduates (and others) possess an adequate vocabulary for this purpose. In discussing the fact that PR subjects were significantly more likely to say that an A statement referring to an empty class of cards is false I mentioned the possibility that they might mean by 'false' simply that there was something wrong with the statement or, more specifically, that it 'implied' something which was not the case (viz., that there were cards of the kind in question). An uncertainty about the appropriate use of the basic terms 'true' and 'false', 'valid' and 'invalid' is apparent also in some of the papers reviewed in the opening chapter of this thesis, for subjects were sometimes asked - for example, in the Thistlethwaite study (1950) - to indicate whether a conclusion followed from the premises (that is, whether the argument was valid or not) by circling a T (for 'true') or an F (for 'false'). Either the psychologist himself did not recognise the importance of not confusing the truth-value of the conclusion with the validity or otherwise of the argument or else he assumed that his subjects would not recognise it.

Closely connected with this issue is the distinction between the case in which a conclusion is established by means of an argument (the argument is valid and the premises are true) and the case where only the first of these conditions obtains. If we ask a group of undergraduates to say simply what is wrong with an argument (which is in fact invalid) what proportion of them will consider only the validity of the argument, what proportion both the validity and the truth of the premises, and what proportion only the truth of the premises? A fragment of evidence relating to this question is presented, in another connexion, in Appendix B (p. . .); research along these lines seems likely to shed light on two of the factors which have been mentioned as sources of error in a syllogistic reasoning task, namely, a failure to adhere to the terms of the task (Henle, 1962, and Richter, 1957) and the tendency to be influenced in one's judgement of the validity of an argument by one's view of the truth or falsity of its conclusion.

I referred, in my discussion of the fact that some highly intelligent persons (and not just undergraduates!) seem to have been prone to the fallacy of supposing that a universal affirmative implies the truth of its converse, to the need for research into the conditions predisposing

a person to make such a mistake, as well as into the development of individual differences of a fairly specific kind in this area in general. With the addition of this dimension to the one just considered one clearly has a programme of research of such proportions that the chief problem, apart from the acquisition of sufficient numbers of subjects prepared to submit themselves to this kind of task, is likely to be one of organisation or, to put it more crudely, of knowing where to begin. It may be the chief virtue of this thesis, at least for the present writer, that it makes this particular problem of decision rather easier than it might otherwise be.

APPENDIX A

VALENTINE'S REASONING TESTS FOR HIGHER LEVELS OF INTELLIGENCE:

A CRITICAL APPRAISAL

Summary This appendix presents evidence bearing on the adequacy of Valentine's Reasoning Tests for Higher Levels of Intelligence, scores on Section B of this test, in the published form and in a form devised by the present writer, having been used as a basis for selecting subjects of widely differing degrees of deductive reasoning ability for the experimental work described in the body of this thesis. A scrutiny of the test's items, suggested by the results of an item analysis using Anstey's d method, reveals that three of them are defective. Doubts are expressed about the way in which the guessing correction in Section B works, about the adequacy of the instructions in Section B and of the time allowed for the whole test, and about the convenience of the arrangements for scoring the test. The predictive validity of the test appears to be at least not greater than that of Terman's Concept Mastery Test, at any rate for the Glasgow undergraduate population sampled. The internal consistency of the test is remarkably high in view of the small number of items, but its alternative form reliability is much lower, there being evidence of a strong practice effect, at least over a period as short as a week. The students who had sat the Terman test as well as the Valentine were on the whole better disposed towards the latter, more of them regarding it as a good test of all-round intellectual ability and as a measure of an ability essential or desirable for the Honours course of their choice. The appendix concludes with the suggestion that while a technically adequate version of the Valentine test could probably not serve as a satisfactory measure of general intelligence because of the degree of its dependence on the special ability to reason logically, and while it would, for the same reason, be unfair to some applicants for admission to university, it is a test without exact counterpart within the field of reasoning ability testing and one which would lend itself to development in a number of directions. It is admitted that if the defects of the test had been apparent before the selection of subjects for the experimental research reported in this thesis had been completed, it would have been advisable to make some allowance for them. As it was, the results of the experiments described in the main part of this thesis make it clear that the two groups selected were genuinely different in the relevant respect.

A.1 Introduction The research described in this thesis was originally designed to serve two related purposes: (1) to identify undergraduates of above average academic ability and attainment who have unusual difficulty with a syllogistic reasoning task and, by comparing them with other students of comparable ability who have much more success with such tasks, to try to discover some of the sources of their difficulty; (2) to add to the resources currently available for the purpose of distinguishing the two types of student. The first, and primary, purpose was, of course, served by the experimental work described in the body of this thesis;

this appendix relates the, rather unexpected, outcome of my attempt to achieve the second.

For reasons set out in chapter 2 I came to believe that the best published measure of performance in the kind of syllogistic reasoning task I was interested in was Valentine's Reasoning Tests for Higher Levels of Intelligence¹ (Valentine, 1954, 1961) and in particular the second, main, section of that test. It seemed to me that for many experimental purposes - most obviously, of course, studies of the effects of 'remedial measures'² - this test, which I shall refer to as the VRT for short, was very much less useful than it might be because available in only one form. The way in which I intended to achieve the second of the aims described above was, therefore, by devising an alternative form of the test. The undertaking seemed to be likely to be a relatively easy one in view of the fact that reasoning tasks, and especially deductive reasoning tasks, have a logical 'form' which can be clothed in a number of different 'contents'. Providing certain precautions suggested by the Wilkins monograph (1928), reviewed at some length in chapter 1, were taken, it seemed reasonable to expect to be able to produce an alternative form unusually close in its parallel to the original.

If this intention had been fulfilled, this appendix would have been devoted to a discussion of the alternative form and to the presentation of evidence of its adequacy. In fact, although an alternative form was devised³ and tried out, as we shall see, on a sample of almost 320 Glasgow undergraduates, the closer acquaintance with the VRT which the preparation of the relevant evidence entailed has led me to have serious doubts about the technical adequacy of the original. As dubious features of the original (Form V, as I came to call it) are embodied in the parallel form ('Form W'), I propose to refer to the latter only in the course of a discussion of the former. This discussion will show, I think, that the VRT as published needs to be altered, at least in some details.

The improvement which would be effected by the appropriate changes would be important, not only for the special experimental purposes mentioned above but also because the VRT was intended by its author as a

1 Reproduced, for convenience of reference, in Appendix B.

2 See the papers by Elton (1965) and Backhouse (1967). Backhouse also produced an alternative form of the VRT which has not been published.

3 Also in Appendix B.

test of general intellectual ability and, as such, as an instrument for selecting applicants for places in universities and training colleges. It has been recommended by Elton (1965) for the assessment of students wishing to transfer from one university to another and, potentially more important, at least in this country, commended by Anstey (1966) as 'an extremely clever and promising test' of high-grade intelligence. Although I shall argue against these uses of the VRT, it seemed useful to take the opportunity which presented itself of adding to the data included in Valentine's own publications results from my own sample about the extent to which performance on the VRT is related (a) to academic attainment, (b) to performance on another, more traditional, group test of intelligence intended for subjects of high ability, Terman's Concept Mastery Test (Terman, 1956) and (c) to performance on my alternative form. I shall also be able to report on the extent to which my own subjects regarded the Valentine test as an acceptable means of assessing all-round intellectual ability or of selecting applicants for university places and I shall conclude with some more general remarks about the plausibility of Valentine's claim that a test of reasoning ability of this kind could serve as a satisfactory measure either of general intelligence or of 'scholastic aptitude'. Amongst the evidence adduced in support of a negative verdict will be the case of subjects such as the PR group whose poor performance on Section B of the VRT served as the starting-point for the research described in the main part of this thesis.

A.2 The Valentine Reasoning Tests: general description The VRT, as I have mentioned, has two parts, Section A, which consists of four problems which call for 'inductive reasoning', and Section B, which comprises twelve problems in which the task is to evaluate arguments ('deductive reasoning'). Valentine (1954) says the first, third and fourth problems of A require the subject to deal, respectively, with relations of time, space and number. The second problem, it seems to me, involves the application of some of the principles of induction first formulated by Mill (1843). In Section B the subject has to say, of the conclusions of twelve arguments, whether they follow from the premises and, if they do not, to select from three (in the first case) or (in all other cases) four reasons offered the one which 'gives the best reason why the conclusion does not follow from the premises'. In fact there are four conclusions which do follow from the premises and, therefore, eight which do not.

It will be clear from the above that the items of Section B are

of the choice-response type. In Section A the third and fourth items are of the inventive-response variety: the subject simply says (in item 3) whether D is nearer to A or to C or (in item 4) on what principles the father's money was distributed. The first and second items are somewhat unusual in the sense that the subject is asked to express his conclusion, about the murderer or about the source of the poison, by underlining, scoring out, or putting a question-mark against, various statements relating to his conclusion, according as he regards them as necessarily true, necessarily false, or else doubtful. Correctly applied, this system should express different conclusions via characteristically different patterns of response, although in scoring these items the response to each statement is treated as if it were independent of all the others.

Form W was constructed along exactly similar lines. In Section A the problems were once again couched in terms of time, space and number relations, with item 2 involving something like the inductive methods of Mill. An attempt was made to retain the special features of the original problems, for example, the notion of a locus of points in the solution to item 3 and the double principle in item 4. The solutions to all problems except the first were, however, different from the corresponding V solutions (as I think subjects would have assumed, particularly in view of the special instructions¹ issued on the occasion of the second VRT testing session). In Section B the 'logical form' of the arguments and of the reasons offered was exactly as it had been in the original form while the content or subject-matter of the items was changed.² To prevent subjects simply recalling their responses in the form they had completed first and thus spuriously inflating the correlation between scores on the two forms (Anastasi, 1968, p. 80), I changed the order of the items apart from the first, 5, by far the easiest, 11, and the last, double, item, 15A and B, the most difficult. In all cases, moreover, the order in which the reasons were presented was different in the two forms.³ One other precaution, suggested by the work of Wilkins (1928) and Janis and Frick (1943) discussed in chapter 1, was to try to equate the two forms with respect to the (perceived) truth-value of their conclusions.

1 See Appendix B.

2 I have to thank Mr Eric Toms of the Logic Department of Glasgow University for independent confirmation of the logical equivalence of the items of Section B in the two forms.

3 A comparison of the two forms and of the correct solutions to them (Appendix B) will make the extent of the differences clear.

The Valentine test has a time limit of 55 minutes, subjects being warned after 20 minutes and advised to proceed to Section B.

A.3 Testing arrangements, means and standard deviations The members of the entire Ordinary Psychology class at the University of Glasgow in 1967-8 took the two Forms of the Valentine on successive Fridays and Terman's Concept Mastery Test on the intervening Tuesday.¹ For the Valentine tests the class was divided into two halves, those whose surnames began with the letters A to L doing Form W, and the other half Form V, first. The total number of students in the class was 325. Of these 322 sat the CMT, 320 Form W and 318 Form V. Of the 158 who sat Form V first, the group in which we shall be primarily interested in subsequent sections, 31 were male Arts students, 76 female Arts, 38 male Science and 13 female Science students. All distributions are set out in Appendix B; Table 1.1 presents means and standard deviations for various groups on the three tests.

TABLE A.1

MEANS AND STANDARD DEVIATIONS OF VARIOUS GROUPS ON THE TWO FORMS OF THE VRT AND ON THE CONCEPT MASTERY TEST (CMT)

		N	Mean	S.D.
VRT Form V	A-L subjects	160	43.2	11.6
	M-Y subjects	158	36.3	13.9
VRT Form W	A-L subjects	161	39.1	12.1
	M-Y subjects	159	45.0	12.2
	All subjects	322	104.2	27.4
CMT	Arts subjects	218	108.1	26.7
	Science subjects	104	96.2	29.0

It will be seen that the two Forms are reasonably close in terms of the above statistics - as indeed they are in terms of the underlying distributions (Appendix B) - although Form W proved to be slightly easier - for reasons which remain obscure. The mean for my M-Y group is slightly higher than the value quoted by Valentine for undergraduates drawn from

¹ Special thanks are due to Professor R. W. Pickford for permission to carry out this testing program, without which none of the research described in this thesis would have been possible.

a variety of English Universities, namely, 34.8. In the absence of a better understanding of the factors which promote success in reasoning tasks, especially of the Section B variety, this small difference may perhaps be attributed to the fact that almost half of the Arts students in my sample had taken a philosophy class in their degree and that 20 per cent said that they had been helped by this in their answers to the VRT.¹ On the other hand, we have no comparable information about the students who made up Valentine's sample and the difference is not a significant one. I have given separate Arts and Science results for the CMT because of the large difference in favour of Arts students on this test, a difference which appears to be due to a bias in this test towards subjects with a particular kind of educational background and one, therefore, which I have had to take account of in the discussions of later sections of this chapter and in the selection of subjects for the experimental studies described previously. There was a difference between faculties in the opposite direction in the case of the VRT and also a difference favouring men rather than women. Neither difference was significant, however, although the difference between male Science and female Arts students, the extreme cases, was significant.

A.4 Item difficulty and validity The Valentine does not lend itself easily to item analysis inasmuch as only four of its items appear to be of the usual pass-fail variety. It is true that Anstey's 'E' indices (Anstey, 1966, p. 103 f.) are supposed to deal with cases in which items are not of this kind. However, the difficulty which has to be dealt with in the VRT does not concern items which subjects have omitted or failed to reach (as the E indices appear to do) but arises because of the different degrees of success recognised in each item. This is most obvious in the case of the four items of Section A where the possible range of scores is, respectively, 0-6, 0-8, 0-4 and 0-5.² It also holds for the eight items of Section B where the conclusion does not follow from the premises: a subject scores 5 points for complete success, 0 for a correct verdict about the validity of the argument but without the correct reason, and -1 for a wrong verdict.

1 See responses to the questionnaire described in section A.7 and reproduced in Appendix B.

2 It may look at first glance as if the items 1 and 2 are really composed of, respectively, six and eight items of the pass-fail variety. In fact this is not the true state of affairs for, as I have mentioned in A.2, the various items are not independent: a subject who concludes that Frogger was the murderer, for example, will produce a pattern of responses which reflects this fact and his score is really a function of two things, his ability to draw the correct conclusion and to 'express' it correctly.

Valentine himself compared the success rates of the top and bottom thirds of his largest single standardisation sample, 92 male training college students. Instead of the more usual comparison between the proportions of each group passing the item, the special nature of the scoring system just referred to must have obliged him to compare the total or average scores of the two groups on each item. He reports that all items had a satisfactory 'validity score' on this criterion except 15A which, he says, was too difficult even for the ablest members of this group. In view of the small number of items in the test, and of the fact that it is intended for use in the selection of candidates for training college, as well as university places, it may be thought that this result of Valentine's own item analysis should have suggested at least a modification of the item in question. As for the relative difficulty of items, Valentine must again presumably have judged this, not by the proportion passing but by the percentage of possible score for each item actually achieved by the group (the method I adopt for Section A). The results of this part of Valentine's analysis seem to have justified the order in which the items are arranged at least to the extent that the last item in Section A and the last two or three items of Section B proved to be the most difficult.

I decided to attempt a different approach to the problem of item analysis, at least so far as Section B is concerned. Even in Section A it is clearly possible in principle to dichotomise responses to an item as correct or otherwise (pass or fail) but there can be no doubt that this produces anomalies: on this criterion, for example, item 2 is the most difficult because relatively few subjects score full marks on it; on the other hand, many subjects lose only 1, 2 or 3 points (out of a possible 8) and very few indeed score 0 - whereas 0 is a relatively common score on items 3 and 4. In Section B, on the other hand, the decision to treat the items as if they were pass-fail in character is a good deal more easily justified if only because, as we have seen, four of the items are in effect of this form anyway. The others can be so regarded if we cease to make a distinction between a correct verdict with a wrong reason and a wrong verdict. In terms of the subject's response this approach ceases to regard the task as two part in character - first deciding whether the conclusion follows from the premises and then, if it does not, identifying the reason which best explains why it does not. Instead, items are treated as if they were of the familiar choice-response variety, the subject's choice being between 'Yes', 'No (i)', 'No (ii)', 'No (iii)' and, in all items except 5, 'No (iv)'. It is not a view which one readily accepts for it

is a common enough experience that one suspects a flaw in an argument before one can say what it is: one feels that the distinction between correctly judging of the validity of an argument and recognising a statement of the fallacy, if any, it contains is one which should be preserved. In a test construction context, however, there is the difficulty, obviously felt by Valentine, of distinguishing between a subject who gets the correct answer by chance and the subject who genuinely does know that an argument is faulty but cannot identify the reason. It is a problem which exists, apparently in unmitigated form, in connexion with the four items where the conclusion does follow from the premises and in the next section I shall discuss ways of coping with the problem. In the meantime, it may be enough to support the decision to regard the items of Section B as if the subject did in every case choose between four or five alternatives from the outset by reference to the answers of subjects to item 35 in the questionnaire referred to above: only 21 per cent of Arts, and 28 per cent of Science, subjects said that they never looked at the reasons before deciding about the validity of the argument. To an unknown extent, therefore, it seems that the view I am taking of the items of Section B corresponds with the view taken by subjects and, amongst other things, this may explain why the items in which the conclusion does follow and where the chances of guessing the correct answer are in theory even turn out to be, on Valentine's analysis as well as on my own, as we shall see, to be at least as 'valid' as the items in which the chances of a correct guess are in theory much less.

A disadvantage of treating the items of Sections A and B differently is that it is not possible to compare them in respect of difficulty and validity. However, this is not, I think, a serious disadvantage, partly because it is obviously appropriate that items increase in difficulty within each section and partly because the relationship between these two types of item and total score must be expected to favour the items of Section B, which contribute twice as much to the total.

Section A Table A.2 presents the percentage of the maximum possible score actually obtained in the items of Section A by members of the M-Y and A-L groups and also the product-moment correlations between item score and total score. I have included data from both groups in this and other tables in the present section because of the unmistakable evidence they generally present of the effects of the A-L group's exposure to Form W.

TABLE A.2

INDICES OF THE DIFFICULTY AND VALIDITY OF ITEMS IN SECTION A OF THE VRT

<u>Item:</u>	<u>M-Y subjects*</u>				<u>A-L subjects</u>			
	1	2	3	4	1	2	3	4
Percentage of max. possible score actually obtained	79.0	67.2	61.4	40.7	87.5	71.3	50.9	45.7
Product-moment correlation with total score	0.43	0.35	0.29	0.37	0.22	0.37	0.56	0.27

* This group did Form V first

All r's are significant. Their different sizes must to some extent reflect the different contributions the different items make to total score - item 2 most and item 3 least. If we take this into account, it seems especially safe to conclude that there is little difference between the items in respect of validity, a result which I find surprising in view of what appears to be the superiority of the third and fourth items. These are not only in the preferred inventive-response form but also call for an effort of concentration of the kind required for success in the items of Section B; the first and second items, in contrast, are relatively simple and the exact form in which the subject is asked to respond is, to my mind, rather dubious.¹ As regards difficulty, the order in which they are presented appears to be the correct one. One curious aspect of the results for the two groups - curious in that it differs from all but one of the other 15 items of the test - is that the difficulty of item 3 seems to have been greater for the A-L group than for the M-Y group. This may be thought to be a case of negative transfer, for the solution to item 3 of Form W permits of a definite answer to the question about the distance of D from A and C whereas item 3 of the original Form does not. The improvement on item 4 apparently produced by practice on Form W may be

¹ As one subject remarked in response to item 11 in the questionnaire, there is something odd about having to say (in response to item 1) whether it must be true, or must be false, that Frogger could not have been the murderer, or that it may be true that he could have been. The objections in the case of item 2 are less serious and the point of asking for the particular kind of response in question is clearer in this case. On the other hand, a subject who assumes that there is only one poisoned food, and consistently expresses his conclusion that it must therefore be the fish or the cheese (but not both) in terms of the eight statements scores one point less than a subject who makes the more serious mistake of accepting that there may be more than one poisoned food but denying that the food which only one of the victims ate could be poisoned.

explained mainly in terms of the number of subjects having time to attempt it: there were only 18 members of the A-L group who showed no sign of having made a beginning with this item (by leaving 'working', for example) as compared with 33 in the M-Y group. In other words, practice probably did not so much improve performance on item 4 specifically as, by speeding responses to earlier items, improve the circumstances in which item 4 was attempted.

Section B For Section B I have carried out a full-scale analysis according to Anstey's d method (Anstey, 1966, chap. 9). The full results are presented in Appendix B. Table A.3 presents d and D , together with the 'index of easiness', $I(E)$. $I(E)$ is a figure which attempts to allow for the fact that not all those who do not attempt an item would necessarily have got it wrong - or would necessarily have got it right. It is a compromise between $100R/N$ (where R is the number of subjects who got the item right) and $100R/T$ (where T is the number who reached the item), the former tending to exaggerate the difficulty, and the latter the easiness, of an item which all subjects did not reach as compared with an item which every subject did reach. The formula for $I(E)$ is: $I(E) = 50(R/N + R/T)$. The index d is obtained by subtracting the mean score, for the test as a whole, of those who did not get the item right, $m(W)$, from the mean score of those who did, $m(R)$. D is found by dividing d by the square root of the unbiased estimate of the variance of the population, in the case of the M-Y and A-L groups in fact equivalent to the standard deviation, at least to two decimal places. D is a particularly useful index for present purposes inasmuch as Anstey (ibid., p. 133) is able to distil his experience as a test constructor in terms of it, claiming that if $D \geq 1$ the item is highly satisfactory, if $\frac{3}{4} < D < 1$ the item is satisfactory, if $\frac{1}{2} < D < \frac{3}{4}$ the item is only fairly satisfactory, if $\frac{1}{4} < D < \frac{1}{2}$ the item is dubious, and if $D < \frac{1}{4}$ the item should be scrapped. (All of this on the assumption that the item is not unsatisfactory on other, perhaps logical, grounds.) It also makes it legitimate to compare the results for my two groups, for whom, as we have seen, the standard deviations differ.

It is apparent from Table A.3A that the order of difficulty of the items in Section B is less satisfactory than in Section A. In particular, items 8 and 11 come too early and items 9 and 13 too late. It is also to be noted that the level of difficulty of the last two items (which, as we shall see, there are other grounds for regarding as unsatisfactory) falls short of the point at which it is as easy to choose the correct

TABLE A.3

I(E), d AND D FOR THE ITEMS OF VRT SECTION B

A M-Y subjects (N=158)

Item:	5	6	7	8	9	10	11	12	13	14	15A	15B
I(E)	91.1	60.1	62.0	58.9	74.1	63.3	58.2	65.2	73.4	29.8	18.3	18.9
d	17.3	13.5	8.3	9.0	16.3	15.6	17.8	13.0	16.2	13.1	12.1	12.3
D	1.24	0.97	0.59	0.64	1.17	1.11	1.28	0.93	1.16	0.94	0.86	0.88

B A-L subjects (N=160)

Item:	5	6	7	8	9	10	11	12	13	14	15A	15B
I(E)	98.8	76.6	60.8	76.9	91.3	72.5	74.4	79.4	83.6	32.7	29.0	36.1
d	16.1	9.7	7.0	12.1	9.4	11.8	12.2	12.2	11.1	11.9	10.2	10.1
D	1.39	0.83	0.60	1.04	0.81	1.02	1.05	1.05	0.95	1.02	0.88	0.87

answer by pure luck as it is by serious effort (Anstey, *ibid.*, p.216). With groups of lower ability - such as the applicants for places at university and, especially, training college for which the test was intended - the outcome would presumably be rather worse. As between the groups, practice on Form W produces gains in rates of success in all cases except item 7 where the rate actually falls. Gains on other items range from the substantial to the enormous with item 14 the only exception, the gain in this case being only 3 percent. The average level of difficulty of this section (and, as it happens, of Section A) appears to be on the low side and would be more so if items 15A and 15B were replaced by easier ones. On the other hand, the optimal level of difficulty depends on the purpose for which the test is to be used and a test which is to serve as a means of selecting candidates for places in a university - not to mention a training college - might be expected to be on the easy side for subjects who have actually passed into university.

Turning now to the validity of the items, Valentine's finding that the 'Yes' items, 6, 9, 12 and 13, are as valid as any, despite their apparently greater susceptibility to guessing, (Valentine, 1954, p. 27), is confirmed in my own analysis, at least so far as the critical M-Y group is concerned, the average D for these items being 1.06 as compared with 0.95 for the other items. (For the A-L group the figures are 0.91 and 0.97

respectively.) This result has already been referred to in this section, it being suggested that subjects may fail to decide on the validity of an argument before inspecting the various reasons why it may not follow and so reduce their chances of guessing right on items in which these would otherwise be 50 per cent. In terms of Anstey's classification of items on the basis of their D values all except 7 and 8 are clearly satisfactory on the results for my M-Y group, 7 being doubtful also with the A-L subjects.

The advantage of the full-scale d method of item analysis is that it focuses attention on some of the finer details of an item and, in particular, may provide evidence which supports doubts engendered by a lowish d, as well as drawing attention to items which might otherwise have passed without serious question. More specifically, Anstey suggests that items in which the mean score of subjects putting an erroneous response approaches the mean score for subjects putting the right answer call for further investigation. It is a little difficult to be sure how close the approach has to be before action is called for. Inspection of the full record of the item analysis of Section B for the M-Y group (Appendix B) suggests that item 7 may be such a case, for here five subjects with a mean score of 38.0 (as compared with a $m(R)$ of 39.3) have omitted to make a choice from amongst the reasons offered. Doubts about item 8 are also encouraged by the fact that the 24 subjects who thought the conclusion does follow from the premises have a mean only 4.3 below $m(R)$ - as well as by the fact that two competent subjects omitted the item. Doubts are raised about 15A by the high means of subjects who chose 'No (i)' or who failed to choose a reason. Corroboration of all these points is to be found in the analysis of the responses of A-L subjects.

Having looked again closely at these three items I must confess to finding no logical fault with 8 and, in particular, no justification for the view that the argument is valid. The mistake is an interesting one, however, because it may be due to a kind of misunderstanding which has already been the subject of much comment in the main part of this thesis: the argument would be valid if - and, as reason (iv) points out, only if - 'always' could legitimately be construed to include 'only' as well, that is to say, if the statement, 'All fools make this mistake' could be taken to mean that only fools do so - or that all those who make this mistake are fools. We have already noted (p. 145) that people of the highest ability are susceptible to this misunderstanding.

Scrutiny of items 7 and 15A, on the other hand, produces a different outcome and suggests that they must at least be altered. In the case of 7 the difficulty is that none of the reasons offered (including, of course, the 'correct' reason) is logically satisfactory. To be specific, reason (ii), the one supposed to be correct, is an argument incorporating a fallacy almost exactly comparable with the fallacy which mars reason (iii). The fallacy is apparent if the reason is restated as follows: 'If some ten-pound householders do not vote for Mr B (as the premises allow) and if all those who vote for Mr B are Whigs, it follows that some householders are not Whigs.' In formal terms this is an example of the illicit process of the major term and, oddly enough, it would elude someone who made the mistake referred to in the previous paragraph, for the conclusion would follow from the premises stated plus the converse of the second one, 'All Whigs vote for Mr B.' In point of fact^{it} is the fallacy involved in the original argument (which would be valid if the converse of the second premise could be assumed to be true), and the present item would, therefore, appear to be a good example of a test constructor being hoist with his own petard! The item would escape criticism if one of the reasons simply pointed out that 'All who vote for Mr B. are ten-pound householders' does not imply that all ten-pound householders vote for Mr B - and that the conclusion follows from the premises only if the latter is assumed to be true.

The fault in item 15A (which it shares with 15B) is a different one and less easily remedied. Before I attempt to set it out, it is perhaps necessary to say a word in defence of a procedure which will have as its conclusion the suggestion that not only 15A, which showed up in the item analysis as somewhat doubtful, but also 15B, of which the same could not be said, needs to be altered to a marked degree. The point simply is, as Anstey remarks (*ibid.*, p. 78), that technical adequacy, as indicated by the results of an item analysis, is a necessary, but not a sufficient condition of an item's being retained unchanged in a test. 'If an item has a logical flaw or other real weakness,' he says, 'it should not be used, however favourable the evidence from item analysis. To be worth using, an item must be known to work well in practice, but it must also be sound and defensible against theoretical attack.' I shall try to show that this is not true of 15A and 15B in their present form.

The flaw in these items may be described as being logical in character or as consisting of a kind of 'catch'; in either case the test

constructor may, once again, have been an unwitting victim of his own ingenuity. Throughout Section B the subject's task has been to say whether the conclusion of an argument follows from the premises. In 15A the 'conclusion' is of the form, 'It would be incorrect to conclude that.....' and in 15B, 'It would be wrong to conclude that'. Obviously, both these 'conclusions' are themselves verdicts about conclusions - to the effect that the relevant conclusion does not follow from the premises - so that in effect the subject is being asked to accept or reject a verdict of the kind he has previously been asked to reach himself. Because this verdict is an unfavourable one, accepting it - saying 'Yes' in these items - is tantamount to saying 'No' ('the conclusion does not follow') on previous items and, of course, vice versa. The result is that many subjects, for the first time in the test, respond by underlining 'Yes' and then putting a cross against one of the reasons. (35 out of the 158 M-Y subjects did this.) More important, perhaps - after all, the subjects just mentioned had got the answer wrong - other subjects may very possibly have underlined 'No' and then assumed that there was no need to look for a reason, an assumption reinforced, as we shall see, by the character of the reasons themselves. It is not possible to be certain how many subjects fall into this second category because, of course, some of the 24 subjects who omitted to put a cross against any of the reasons may simply not have been able to decide amongst the alternatives.

There is an aspect of the alternatives of 15A and 15B which may suggest that even Valentine was not clear about the extent to which the task had been changed in these last items and which in any case makes these alternatives highly unsatisfactory. Supposing the subject's perspicacity to be of such an order that he recognises that in saying 'No, it does not follow that it would be incorrect to conclude that ... ', he would be saying that the conclusion does follow from the premises, and supposing that he also realises that a 'reason' in the present instance must therefore be something which shows that the conclusion does follow, then he should have no difficulty in selecting the correct alternative, for three of the four reasons offered claim to find a fault in the argument and only one even looks as if it proved the argument valid. Consequently, subjects who underline 'No' and choose one of the wrong alternatives may be suspected of meaning 'Yes, it would be incorrect ...'. Small wonder that Hallworth (1963) in his factor analysis of the test found that items 15A and 15B belonged in a class of their own - though he attributed their special position in the test to their being of a 'different order of diff-

iculty' in the sense that they called for an ability to deal with 'more complicated' reasoning problems. I think I have said enough to suggest that Hallworth's remarks are true, if not quite in the sense presumably intended. In my view, these items must be changed, preferably by dropping the verdict-giving form of the 'conclusion', but at least in the respect that all the distracters at least appear to establish the validity of the argument. Undoubtedly the more radical change would make the test easier but there are other ways of increasing its difficulty - for example, by making the arguments more complex or by refining the differences between the alternative reasons from which the subject has to choose.

A.5 Other specific points of criticism (a) The guessing correction in Section B I have already mentioned the scoring system in Section B, where 5 points are awarded for a correct 'No' plus the appropriate reason, 2 points for a correct 'Yes', 0 for a correct 'No' without the correct reason, and -1 for a wrong 'Yes' or 'No'. The last two scores represent a correction for guessing which, as Anstey remarks (*ibid.*, p.231), gives 'rough justice'. There are two respects in which the justice seems particularly rough as compared with guessing corrections in general. The first is that a double penalty for guessing operates in the case of 'No' items: if a subject guesses wrong, he scores -1 and if he guesses right (or, more strictly, is deemed to have guessed the right answer because he fails to identify the correct reason for his verdict) he scores 0. (On the other hand, while a wrong guess in a 'Yes' item scores -1, a correct guess scores 2.) Secondly, there is an arbitrariness about the way in which the resulting penalty points enter into the total score for the test. So long as a subject's score for Section B is not less than zero, all penalties are deducted from his total score. If, on the other hand, his Section B score is a minus quantity (it was actually as low as -6 for one of my subjects), the points below zero are not deducted from his total score. The justification Valentine offers for this procedure is that 'marks are deducted merely to discourage students from guessing' (1954, p. 33). It is, however, the threat of deduction and the belief, on the part of subjects, that the threat will be carried out, not its actual execution, which discourages guessing - if anything does. The question of how it should be carried out in detail must be decided on other grounds, and I can think of none which justifies the arbitrary procedure suggested by Valentine.

I do not, of course, dispute that there is a case for a guessing correction in this section of the VRT. If subjects obey the instructions

and begin by choosing between 'Yes' and 'No' and only later, and in the second instance, look for the reason, then the chances of guessing the correct answer at this stage - the only stage in the four 'Yes' items - are obviously 50-50. I suppose it might further be argued that the double penalty in 'No' items is justified simply in terms of its practicality: in these items we have evidence of guessing (failure to choose the correct reason) which is not available in the case of the other type of item. The weakness of this second point is, of course, that it implies that a correct 'No' can safely be regarded as the product of logical reasoning ability, as opposed to pure luck or intelligent guesswork, only if it is supported by other evidence (the ability to recognise a statement of the fallacy in question). What is true of a 'No' response, however, must surely be true of a 'Yes' response as well, so that a correct response in the latter case must be regarded as just as much or as little the product of guesswork as a correct response in the former. The implication of this clearly is that correct verdicts in the two kinds of case deserve to be treated in the same way: either a correct 'No' deserves to score 2 points or a correct 'Yes' deserves to score none. Which of these alternatives we prefer will depend on how we evaluate the part played by guessing in this section of the VRT. If we think its role is a relatively minor one - as, perhaps, the high validity of the 'Yes' items would suggest - we should opt for the first; if we think its probable role a major one, as on paper it may be, we should prefer the second - or, arguably, the third, and more radical, alternative of dropping 'Yes' items altogether and giving no credit for correct 'No's' unsupported by the correct reason.

Guessing corrections in general have been the source of a great deal of controversy, it being a matter of uncertainty as to how fairly they work. The preferences of Guilford (1965) and Anstey (ibid.) are clearly for eliminating the need for such corrections by altering the items in a test in such a way as to reduce the likelihood of success by chance. Thus Anstey regards the choice-response type of item as the least suitable for high-grade tests because it limits the difficulty of items: if they are passed by fewer than $1/x$ th of the population for which the test is intended (where x is the number of alternatives from which the subject chooses) a subject is more likely to get the right answer by chance than by reflection or skill - even assuming that all x alternatives are equally plausible.¹ The problem is obviously at its greatest

1 It may be apparent from the item analysis schedules (Appendix B) that

where the subject has to choose between only two alternatives, as we have seen he does in the 'Yes' items of Section B if he abides by the instructions. Such items Anstey regards as 'thoroughly objectionable' and in general he would prefer to see choice-response items replaced either by ones in which the subject has to supply the answer ('inventive-response' items) or where he is called to match items from one list with items on another ('matching-response' items) where the probability of chance success is greatly reduced.

It is, in fact, the former of these alternatives which Anstey seems to favour for the VRT which, he says, 'might have been even better if cast entirely in inventive-response form' (ibid., p. 231). My own view is that while this may be entirely feasible in the case of the two items of Section A which are not already of this form, there would be very considerable difficulties in the case of Section B. The inventive-response items which Anstey cites usually call for an answer which is a single word. In such a case the scorer's task is a relatively easy one: he has only to be able to distinguish the correct word from all others. Even in items 3 and 4 of the VRT, where subjects are required to write a phrase or a sentence, some problems arise about the equivalence of different phrases or sentences. In the case where a subject was asked to say what, if anything, was wrong with an argument, the difficulties involved in discriminating right from wrong responses would, I think, be very much greater, for the same point may not only be made in different words but also with different degrees of explicitness.¹

In any case it is not clear how an inventive-response format could be made to cope with the case of valid arguments for here the subject would be required to find a way of restating an argument in terms not identical with those used in the original - and the problem of finding a compromise

this condition does not hold in the case of the VRT Section B: even if we regard the subject's choice as being always between four or five alternatives, it seems that some of these have very little plausibility and so reduce the difficulty of the task facing the intelligent guesser. Valentine says that he devised the reasons himself; it would clearly have been better if they had been based on empirical evidence about the kind of reasons which students find plausible. I have myself made a tentative start with the collection of evidence of this kind in the exercise referred to in Appendix B.

¹ In Appendix B I have set out the explanations given by 45 members of the Ordinary Psychology class of the invalidity of one of the VRT items. I think it will be apparent from this that this type of item could be scored only by someone whose logical ability was at least as great as any of the subjects tested, for the variety of ways in which the same point can be made seems to be indefinitely large.

between something which is too similar to the original and something which is just not similar enough is one which it would be extremely difficult to find an adequate way of explaining, in the first place, and of judging, in the second. I have suggested that the 'Yes' items of the VRT may not be as susceptible to distortion through guessing as they seem at first sight - and as they would be if subjects obeyed instructions. If, in spite of this, we were to regard them as the 'thoroughly objectionable' type of item in which the chances of success by guessing are 50 per cent, it might seem desirable simply to eliminate them from the test. I do/^{not}mean that the task might continue to be represented as one in which the subject has to judge the validity of the argument and then, if it is invalid, choose the reason which best explains why it is so. Early investigations of syllogistic reasoning (for example, that of Morgan and Morton, 1944) did employ tests in which none of the syllogisms permitted one to draw a valid conclusion, but, like Henle and Michael (1956) I think it undesirable that a test should run counter in this way to the natural expectations of the subjects. Of course, given the form of these early tests, it was not possible to explain at the outset that all the items were in fact invalid: the subject's task was simply to say to say which, if any, of the conclusions listed followed from the premises. In a task such as that presented in Section B of the VRT, however, it would be perfectly possible to begin by saying that there was something wrong with all of the following arguments and ask subjects to say which of the reasons offered appeared to them best to explain what this was. When the problem was presented in inventive-response form - as the data from the preliminary study referred to in Appendix B will illustrate - this task proved to be a very difficult one for some subjects and revealed wide differences in the adequacy with which it was tackled. In choice-response form it should be possible to vary the difficulty of the task by increasing or reducing the differences between the various reasons offered.

Such an arrangement would have the apparent disadvantage that it would fail to involve what is presumably the primary ability of distinguishing valid from invalid arguments. It might therefore be suggested that a superior solution to the problem of the 'Yes' type of item - if it is a problem - would be to devise a task which would play a similar role in this type of item to the role played by the 'reasons' of the Valentine test in cases where the conclusion does not follow from the premises. The nature of such a task may already be clear from the defective items 15A and 15B, as well as from the discussion of the inventive-response alternative.

It would consist of the identification of an adequate restatement of the original argument, and, once again, the level of difficulty of the task could be varied by making the alternatives from which subjects had to choose more or less similar.

The chief disadvantage of this second alternative - apart from the very considerable effort involved in producing a satisfactory version of it-would be likely to be its length. It seems probable - though the responses of some subjects to 15A and 15B might suggest otherwise - that the distinction between reasons which purported to show why a conclusion did not follow from the premises of an argument and reasons purporting to show that it did is one which subjects would find it easy to make. In that case, four or five reasons of each kind would have to be offered in order to reduce the chances of success by guessing to acceptable proportions. On the other hand, the sheer effort involved in reading so many alternatives could be used to encourage subjects to begin by deciding whether the argument was valid or not if the two kinds of reason were presented in separate sets. In these circumstances, a subject who reached the right verdict could save himself considerable time and trouble by limiting his attention to the alternatives of the appropriate set.

In summary, then, the guessing correction as at present operated in the VRT Section B appears to be unacceptable for two reasons. If it seems necessary to deal with the case of valid arguments, where the chances of success by guessing are unacceptably high if the subject adheres to the procedure laid down in the instructions to Section B, at least two possibilities suggest themselves. In view of the labour involved, and of the evidence that the items in question are not invalidated by their apparently objectionable form, it might be thought sufficient to modify the present arrangements with regard to the application of the guessing correction - in particular, by treating the penalties which accrue from it in a uniform way for all subjects and by awarding the same score for correct verdicts on invalid arguments as for valid ones - and to attempt to produce distracters of more equal plausibility.

(b) The adequacy of the instructions in Section B In his chapter on high-grade tests of ability Anstey insists on the importance of adequate instructions. It has, in fact, become a commonplace of mental testing that a test should include, not only very explicit instructions, but also, as Anstey (*ibid.*, p. 244) says, 'at least three worked examples of

each type of item, including one that is fairly difficult'. On the first of these counts Section B seems to be reasonably adequate; on the second, on the other hand, it clearly is not. Valentine says (1954, p. 25) that 'the first test of Section B was deliberately made very easy as a kind of practice test to ensure that the method is understood'. In other words, the function of item 5 is partly that which would normally be served by a worked example. Such an arrangement appears to me to be both undesirable in itself and also, in all probability, ineffective: undesirable because a test comprising only 16 items can ill afford to have one of them devoted to a secondary purpose, and ineffective because a subject has no way of knowing if the response he makes to this item is the correct one, that is, if he has 'understood the method'.

In their answers to item 33 of the questionnaire only 9 per cent of subjects said that they had not found it easy to understand what they were supposed to do in Section B, while 30 per cent of Arts and 39 per cent of Science students said they found it very easy to understand this. On the other hand, between a third and a half said, in answer to item 43, that they found the Form they did second easier and by far the commonest reason given for this was that they knew better what they were supposed to do. And, of course, the reality of this phenomenon is attested to by the generally large gains in score, on the original Form of the test, by subjects who had previously done Form W.

It may, of course, be that a practice effect of this kind would survive the introduction of 'three worked examples'. There would also be the problem of explanations to subjects who could not see why the correct answer to an example was correct - though, of course, this would not be a problem peculiar to the VRT Section B. Finally, and probably most serious from a practical point of view, the introduction of worked examples would inevitably mean an increase in the time required for the administration of the test. Nonetheless, the case for devising a better way of introducing the problems of Section B seems to me to be a very strong one.¹

¹ As an example of the way in which this might be done one might cite the Hertzka and Guilford Logical Reasoning test where 'four carefully chosen practice items are provided, since this kind of exercise is new to most individuals' (Hertzka and Guilford, 1955, p.1). The examples are accompanied by attempts to explain why one conclusion rather than another follows from the premises. Other helpful comments are included of which the following is perhaps the best example: 'Notice that a correct conclusion is derived from both statements and from those statements only. A correct conclusion is not just a repetition of the contents of one of the statements. A correct conclusion is not based on other information than that supplied by the given statements.'

(c) The adequacy of the time allowed 26 M-Y subjects made no attempt to answer item 4 (did not even show 'working'), 9 made no attempt at 15A and 9 more no attempt at 15B. Since each of these items contributes 5 points towards the maximum possible score of 71, these figures must, I think, be viewed with some concern - especially as the subjects concerned belonged to a highly selected group. In answer to items 29 - 31 in the questionnaire 50 per cent of Arts and 30 per cent of Science subjects said they could have done better if they had had more time. Most of these said they would have spent it on both sections, the remainder being divided rather evenly between those who said they would have spent it on Section A and those who would have spent it on Section B. Valentine himself suggests that a Principal of a training college might find it advisable to extend the present time limit by 15 minutes with a group of candidates of below average ability. In my own view, an extension of this order would be desirable for all groups of subject. Indeed it is obviously only administrative convenience that requires the application of any time limit at all: other high grade tests, such as Terman's Concept Mastery Test and the very different Advanced Matrices Test of Raven have none. It is perhaps of some significance that 48 per cent of Arts and 44 per cent of Science students said, in answer to item 32 of the questionnaire, that the mere existence of a time limit interfered with their ability to concentrate on the task. Abolition of the time limit might eliminate this source of anxiety without noticeably increasing the time taken by the vast majority of candidates.

(d) Arrangements for scoring In its published version the VRT is difficult to score. Answers are written in the test booklet immediately below the question itself and as there are never more than two items to a page, the use of an aid to scoring, such as a stencil, is impracticable. It is true that with so few items in the test as a whole an experienced scorer could probably memorise the correct responses as well as the scoring system, but the risk of scorer error in such a case is, I think, too obvious to need stressing. The existing arrangement is also, to my mind, a wasteful one: test booklets can be used only once. Both these problems could be dealt with, at least to some extent, by the introduction of separate answer sheets.

It is true that Anstey (ibid., pp. 40-1) argues against the use of separate answer sheets on two grounds, first, that subjects may put their answers in the wrong places, and secondly, that they may mark the

in ways
 test booklets/which may affect - adversely or otherwise - the responses of later subjects unless they are discovered in the course of a laborious scrutiny of the booklets after each testing session. The second point is, of course, an important one: depending, perhaps, on circumstances, it may be a foolish economy to attempt to carry out the necessary examination of test booklets rather than buying new ones for each testing. (Moreover, it may be argued that the mere fact that it is possible to re-use the test booklets may tempt careless testers to do so without first ensuring that they are unmarked.) This last consideration apart, however, Anstey's second point is clearly an argument against the re-using of test booklets and only indirectly one which casts doubt on the advisability of separate answer sheets. As to his first point, which does refer to a disadvantage of separate answer sheets as such, I think it is of somewhat reduced importance in connexion with a test where there are only 16 items in all and only 12 in which the possibility of confusion between answers can seriously be said to exist. In favour of the use of separate answer sheets, on the other hand, is, of course, Anstey's insistence (p. 40) that a satisfactory test is one which it is easy to score reliably.

Accordingly, my subjects were provided with separate answer sheets, of which an example, for Form W, is reproduced in Appendix A. It cannot be claimed that the scoring of Section A was thereby made easier or more reliable, but it was possible to score Section B with a minimum of effort using the three stencils which are also reproduced in Appendix A. (If the guessing correction were dropped, of course, the 'Wrongs' stencil would no longer be necessary, and if correct verdicts in invalid arguments were treated in the same way as in valid ones, as I have suggested they should, a separate 'Yes' stencil would also become redundant.)

A.6 The predictive validity and the reliability of the VRT

(a) Predictive validity This is the aspect of the test on which Valentine lays the heaviest emphasis - and with good reason, since its avowed purpose, as we have seen, is to act as a basis for selecting applicants for places in institutions of higher education and the results he presents are impressive. He is able to report, for example, (1961) that graduates with first class honours (N=40) score very significantly higher than those with second class honours (N=115) and these in turn significantly higher than graduates with third class honours or pass degrees (N=22 and 50 respectively). Since most of my subjects have now graduated, I am able to produce comparable, though much smaller, figures

for the University of Glasgow. One difference between my data and Valentine's which may be worth referring to in passing is that his subjects had already graduated at the time of testing. Table A.4A presents the means and the ranges of scores for five categories of student, those who obtained first, second or third class honours degrees in Arts or Science, those who obtained Ordinary (i.e. general) degrees and those whose performance at university was so poor as to lead to their suspension or consideration for suspension. For purposes of comparison I present in Tables A.4B and A.4C the same statistics as they relate to performance on Terman's Concept Mastery Test and on the Higher Grade of the Scottish Certificate of Education, both of which might naturally be expected to provide some basis for predicting success at university. In all three tables I give results only for the M-Y group since it is, of course, only the members of this group whose performance on the VRT is strictly comparable with that of Valentine's subjects. The N's vary a little from table to table because the relevant scores were not available for all subjects.

The Concept Mastery Test, Form T, (Terman, 1956) is a revised version of the test which its author used in his follow-up study of his gifted group at maturity (Terman and Oden, 1947), revised downwards, it should be said in passing, to make it suitable for testing the spouses of the gifted group. It is much more traditional in character than the VRT, consisting, as it does, of two Parts, Synonyms-Antonyms (115 items) and Analogies (75 items), which appear to test mainly vocabulary and general knowledge.¹ I used it partly because of its historical interest, partly because of its traditional character, and partly because it presents problems of administration and scoring which are minimal: it is untimed and scoring is by means of two stencils. Correlations with grade-point averages of .49 for Stanford University undergraduates and .37 for subjects at the University of California Counselling Center suggest that its predictive validity is about average.² The same appears to be true of its reliability as measured by correlations between scores on the original Form A and Form T, the coefficients ranging from .86 to .94 for intervals between testing

1 It departs from tradition in using only these two types of item instead of the five or six types more commonly used. (Compare Army Alpha and its derivatives, Cattell's scales, the N.I.I.P. Group Test 33, etc.) This reflects Terman's frequently expressed conviction that it is in the sphere of abstract reasoning that excellence of the relevant kind reveals itself most clearly and that abstract thinking depends on the number and variety of concepts at a subject's disposal and his ability to see relationships between them - together with the assumption that these things are closely related to size of vocabulary and amount of general knowledge.

2 See Guilford, 1965, p. 103.

ranging from one day to twelve years.

No comparable data is, of course, available about the reliability and validity of performance on the Higher Grade of the Scottish Certificate of Education as an index of academic potential. On the other hand, this way of assessing probable performance at University has the advantage, as compared with any published intelligence test, that it is actually used in the selection of applicants for places at Scottish universities. This implies, of course, a high degree of confidence in its predictive validity and also, one assumes, in its reliability. And since Valentine presents his test explicitly as a way of remedying the defects inherent in the English counterpart of this Scottish system of selection, it is clearly of some interest to have data in which the two can be directly compared. The Scottish Universities take into account, not only the number of passes obtained on the Higher Grade but also the grade of the pass, there being three grades, A, B and C. I decided to consider only the passes obtained in a subject's fifth year at school, there being too many imponderables about passes obtained in the sixth year. (See footnote 3 on p. 42 above.) I obtained an S.C.E. 'score' by awarding 3, 2, and 1 points for passes on grade A, B and C respectively, a rather crude procedure which has since been adopted by Nisbet and Napier (1970) in their study of student success and failure at university. Distributions of the resulting scores are to be found in Appendix B, Arts and Science students, of course, being treated separately.

TABLE A.4

MEANS AND RANGES OF SUBJECTS WITH VARIOUS OUTCOMES AT UNIVERSITY
ON THE VRT, CMT AND SCE (M-Y SUBJECTS ONLY)

A. VRT

		1st Class Hons.	2nd Class Hons.	3rd Class Hons.	Ord- inary	Sus- pended
<u>Arts</u>	N	1	10	1	69	12
	Mean	(64)	41.5	(13)	32.7	34.2
	Range	-	23-64	-	5-65	17-55
<u>Science</u>	N	2	8	2	28	3
	Mean	50.5	51.6	39.5	38.0	27.3
	Range	42-59	13-70	20-59	19-68	13-37

B. CMT

		1st Class Hons.	2nd Class Hons.	3rd Class Hons.	Ord- inary	Sus- pended
<u>Arts</u>	N	1	10	1	69	12
	Mean	(164)	129.2	(117)	104.1	97.0
	Range	-	91-157	-	62-174	36-159
<u>Science</u>	N	2	9	2	27	3
	Mean	107.0	96.6	76.5	95.5	77.0
	Range	100-114	52-132	32-121	36-152	62-95

C. SCE

<u>Arts</u>	N	1	9	1	67	10
	Mean	(10)	6.0	(8)	6.7	6.5
	Range	-	2-12	-	0-13	1-10
<u>Science</u>	N	2	9	2	26	3
	Mean	15.0	8.3	10.0	7.9	6.7
	Range	-	2-20	7-13	2-12	6-7

I have given ranges rather than standard deviations because so many of the samples are too small for the latter statistic to have any meaning. Indeed, it will be apparent that little significance of any kind can be attached to the results for any of the groups except the Ordinary graduates in both faculties and, with diminishing confidence, Second Class Honours graduates in Arts and Science and suspended students in Arts. Within the limits set by the size of the samples, then, there does appear to be a difference in both faculties between the VRT scores of students who gained second class honours and those who graduated with an Ordinary degree. Although the Scottish Ordinary degree is supposed to be different from, rather than inferior to, an Honours degree, I think it would be widely accepted that the quality of students who graduate with the former is generally lower than the quality of graduates who obtain a second class honours degree. To this extent, then, the above result supports Valentine's claim to have devised a test which discriminates within the student population. Moreover, it seems to be superior to the CMT in this respect, so far as Science students are concerned, and to SCE in both faculties. It may be pointed out, on the other hand, that it fails (as the SCE results also do) to find any difference between Arts students who graduated with an Ordinary degree and those who were suspended - whereas the CMT does appear to have found a difference. Perhaps rather little significance should be attached to this aspect of the results, not only in

view of the sample sizes but also because of the heterogeneous nature of the factors which seem to be responsible for a student's doing badly enough to come up for consideration for suspension.

Evidence bearing on the predictive validity of the VRT drawn from other sources is less favourable. Product-moment correlations between SCE scores and VRT results on the one hand and SCE and CMT on the other proved to be as follows: Arts, .32 and .25, Science, .45 and .50. Moreover, when the 15 subjects who went on to graduate with an Honours degree in psychology were ranked in terms of the numbers of answers in their Finals papers which were agreed by three markers to be first, upper second, lower second, third class or fail and these ranks compared with their rankings on VRT, CMT and first year psychology class examinations, the resulting Spearman's rho's, corrected for tied ranks (Hays, 1963) were, respectively, .74, .85 and .83.

The evidence is, of course, far from conclusive. The higher correlation between CMT and SCE results for Science students is offset by a lower correlation for Arts students, and the Honours class sample is small and the results, therefore, of doubtful reliability. On the other hand, one may perhaps draw the restricted conclusion that the VRT on this showing at least does not appear to be superior to the CMT or to class exam results in its ability to predict outcomes at the Honours level. Such a result is clearly no cause for self-congratulation so far as a defender of the VRT is concerned, not only because of the notorious unreliability of class examinations (correlations between results on the two examinations in any academic year are generally of the order of .5) but also because of the doubtful nature of the CMT, especially as a measure of ability for Science students - as we shall see in the next section.¹

(b) Reliability The VRT must surely be one of the shortest tests, in terms of number of items, currently available. It is surprising, therefore, to find Valentine claiming a split-half reliability coefficient as high as .83, equivalent, he says, to a value well in excess

¹ The CMT is singled out for criticism by Anstey, not only because of the 'thoroughly objectionable' character of its first part (where the chances of success by guessing are even) but also because of its dependence, at least in some of its items, on abstruse and specialised knowledge. The test is not mentioned by name. Anstey (ibid., p.219) simply refers to a discussion of 'a test of high grade intelligence of high repute' by Heim in New Society for 7th February, 1963. However, the two items quoted are 108 in Part I and 66 in Part II of the CMT. They are not untypical.

of .9 'in an unselected population'. Backhouse (1967), in a study of the effect of mathematics teaching on reasoning ability as measured by the VRT, reports a split-half reliability of only .75. I estimated the internal consistency of the test by other means, viz., the 'alpha coefficient' (Nunnally, 1970), a statistic found by applying one of the Kuder-Richardson formulas. This index of internal consistency reliability, which can be shown to be equal to the mean of all split-half coefficients resulting from different splittings of the test (Cronbach, 1951), gives a lower value than the ordinary split-half coefficient unless the test is a highly homogeneous one - which, in view of its two Sections, the VRT cannot really be said to be. In the event, however, the value of the alpha coefficient proved to be .96 for the original form of the Valentine test (M-Y subjects) and .94 for Form W (A-L subjects).

These results are important, not simply for themselves but for the basis they provide for a comparison between the reliability of the test measured in this way and its alternative form reliability, for the former represents a maximum value for the test's reliability (Nunnally, p. 552) compared with which the latter indicates the extent to which performance on the test is affected by changes in content, from form to form, and by the exact time in a subject's life at which the test is taken. Backhouse, in the paper referred to above, found an alternative form reliability of .73, which is close to the split-half coefficient we have seen he found for the original form of the test and identical with the corresponding value for his alternative form. In my own study the correlation between scores on the two forms, averaged over A-L and M-Y groups, was .61. a very much lower figure, obviously, than the alpha coefficients reported in the previous paragraph.

The discrepancy between Backhouse's results and my own may, of course, be due to differences in the adequacy of the alternative forms or to differences in the populations sampled: Backhouse's sixth formers would doubtless be a somewhat more heterogeneous group than my undergraduates and the resulting correlation might be expected to be higher in his case. Of more significance, probably, is the difference in the interval elapsing between the completion of the two forms, for whereas my subjects could be expected to show a strong practice effect after only a week, the same would not be true of Backhouse's subjects, for almost two years elapsed between the first and second testing sessions in their case.

The significance, for the experimenter, of the low alternative-form reliability of the VRT over short periods will be apparent: if one cannot in any case expect a close relationship between scores on two forms of the test taken over such intervals of time, then it will be difficult to attribute changes in the scores unequivocally to the effect of any experimental treatment which may have intervened between the two testings. Clearly an attempt must be made to increase the alternative-form reliability of this test - in the first instance, I think, by improving the instructions and by relaxing the time limit, both of these measures which should help to reduce the initial differences between individuals which practice presumably effects most.

A.7 'Face' validity It is clearly of great social importance for a test which is used to select applicants for places at a university to look as if it is a good way of doing this. The success or failure of an application is too serious a matter for the person's concerned for it to be tolerable that any considerable proportion of those who fail should be able to attribute their failure, with any real degree of plausibility, to the inadequacy of the instrument used for selection purposes. I therefore thought it useful to canvass opinions on this aspect of the VRT - and, for purposes of comparison, the CMT - amongst students who had taken the test. This was achieved by means of the questionnaires already referred to on several occasions in this chapter. These were issued on the third testing session and collected either at that time or at the next meeting of the class. Altogether 300 of the students in question responded to some or, usually, all of the 50 items in the questionnaire. It will be apparent, from previous references to it that its purposes were diverse, and one important one will be apparent only at a later point in this thesis. It is reproduced in full, together with the distribution of responses to each item in Appendix A.

So far as the face validity of the VRT is concerned, subjects were asked to say whether they thought (1) that the VRT was a good test of all-round intellectual ability; (2) if they intended to take an Honours degree, that the VRT tested an ability which is essential or desirable for success in the Honours subject of their choosing; (3) that it would be fair to use the VRT as the sole criterion for admission to the Arts faculty of a university; (4) that it would be fair to use it as the sole criterion for admission to the Science faculty; (5) that it would be fair to use it as the sole criterion for admission to the Honours course, if any,

they hoped to follow. The same questions were asked about the CMT and about the VRT and CMT in combination. Students were also asked whether they thought (6) that scores on 'tests of this sort' - i.e., VRT and CMT - ought to be taken into consideration as well as SCE results in selecting applicants for places at university or (7) as well as results in pre-Honours class exams in selecting applicants for places in an Honours course. They were asked to say (9) what, if anything, was wrong with the VRT and CMT as tests of all-round intellectual ability and finally, because I think this aspect of a test is an important one, (10) whether they found the CMT more or less enjoyable to do than the VRT.

As already mentioned, the details of the distributions of responses to these, and other, items are presented in Appendix B. I have considered the views of students from the two faculties separately, but not the views of students of the two sexes, mainly because it is hardly conceivable that different tests should be used for male and female applicants for university places but quite conceivable, at least in Scotland, that different tests should be used by different faculties. An inspection of the responses of men and women on the first five issues listed above shows, as one might perhaps expect, that men were slightly more favourably disposed towards the VRT than women and that the opposite was true of the CMT. The significance of this point is that the proportion of Arts women in the Ordinary Psychology class was higher than in the faculty as a whole. To some small extent, therefore, the attitudes towards the two tests reported below will show an unrepresentatively pro-CMT, anti-VRT bias on the Arts side.

Bearing this point in mind, then, the responses to the relevant items of the questionnaire can be summarised as follows. While only 30% of Science, and 35% of Arts, students thought the CMT a good test of all-round intellectual ability, the corresponding figures for the VRT were 49% and 39%. Science students clearly felt that the CMT would add nothing to the VRT in this respect; on the other hand, 60% of Arts students thought the two tests combined would make a good test of all-round intellectual ability. Of those students who hoped to take an Honours degree (N=63 for Arts and 36 for Science) 28% of Arts and 50% of Science students thought the VRT tested an ability essential to success in the Honours subject of their choice, 57% and 40% thought the ability in question was desirable but not essential, and 7% and 6% thought it neither. The corresponding figures for CMT are: Arts, 15%, 68% and 12%; Science, 14%, 49% and 34%.

Students of both faculties were overwhelmingly against the use of either test, separately or in combination, as the sole criterion for admission to either faculty, though there is strong support for the consideration of results of tests 'of this kind' as well as passes on SCE, GCE, etc. Much the same is true of their use in selecting applicants for admission to an Honours course, with the exception that enthusiasm for their use in an auxiliary role is somewhat reduced. On the question what was wrong with the tests as measures of all-round intellectual ability the answers were mainly along the lines one might expect: that both were too narrow, that VRT favoured students with experience in, or a penchant towards, logic or mathematics, and that CMT favoured those with a classical education or with a large vocabulary or a large store of general information. 60% of Arts students found the CMT more enjoyable than the VRT while 22% found it less, the corresponding figures for Science students being 28% and 51%. This result shows the familiar, and, in view of the contents of the CMT, perhaps natural, pro-VRT, anti-CMT stance of the Science students as well as the ability, amongst Arts students, to distinguish the palatability of the CMT from its apparent validity. The Arts view is the one I should have expected, for the intellectual effort called for in the VRT is, in my view, very much greater. (Hence Valentine's insistence on the importance of good motivation sitting his test.)

Quite generally, despite the overall superiority of the 'face validity' of the VRT when compared with the CMT, it is clear that it would be quite unacceptable to the vast majority of students of either faculty as the sole criterion for admission either to University or to an Honours class. On the other hand, 'tests of this kind' would be acceptable in an auxiliary role by most students of both faculties for both purposes and, despite his strictures on GCE results and headmasters' reports as bases for selection, Valentine does not appear to have intended his test for use in any more exclusive way.

A.8 Review and discussion On the whole the picture painted of the VRT in this chapter has clearly not been a favourable one. It has been shown to have three defective items out of sixteen, an unsatisfactory arrangement for dealing with guessing, inadequate instructions, too stringent a time limit and a layout which makes the scoring of the test unduly arduous. Its alternative-form reliability, as measured over a period of a week, is low in relation to its internal consistency, and although it discriminates rather well within the undergraduate population, it does not obviously do

this better than the CMT, a test whose deficiencies have been the subject of comment by Heim and Anstey. All in all the implication appears to be that the VRT is in need of fairly extensive revision, preferably, of course, based on extended empirical research into the effects of various modifications to its contents, its scoring system and its time limit. However, the effort involved in such an enterprise would clearly be justified only if the idea of the test, as opposed to its present concrete manifestation, is judged to be a good one. It is to the examination of this question that the remainder of this section is devoted.

The usefulness of the VRT can be discussed with outcomes of three different degrees of generality in mind: it can be discussed as a means of measuring either (a) 'academic potential' or (b) general intellectual ability or (c) inductive and deductive reasoning ability. The three are obviously related: reasoning ability is clearly at least a part of general intellectual ability, just as the latter is clearly at least part of what is required for academic success. Obviously, however, one might regard the VRT as a good way of assessing reasoning ability but not as a good test of all-round ability or, of course, as a good test of the latter but not as a satisfactory way of selecting candidates for places in an institute of higher education - or, indeed, vice versa. The view I shall take is that, in the absence of very definite empirical evidence to the contrary, the VRT is likely, on general theoretical grounds, to be more effective - to possess greater 'predictive validity' - the less general of these three purposes it is used to serve.

Valentine clearly regards his test as a satisfactory way of measuring general intellectual ability, 'g', and academic potential. He explains that he confined its contents to reasoning tasks on the ground that, according to Vernon (1950), 'reasoning ability is largely dependent on g'.¹ On the other hand, all the evidence Valentine produces in support of the validity of his test bears on its relationship with measures of

1 Unless I am mistaken, the view Vernon takes in the pages referred to by Valentine is the converse of the one he needs to justify the use of a reasoning test as a measure of g. Vernon says (p. 55): 'A small reasoning or logic group factor could be isolated from specialised tests, but it would be unlikely to add anything to measures of g, v and n in the prediction of the reasoning ability desirable among secondary pupils or college students'. I take this to mean that a good measure of g, v and n will also be a good test of the kind of reasoning desirable among the subjects in question - and not the reverse.

academic success.¹ However, it is certainly the view of Vernon, in the book to which Valentine refers, that success in school or college depends, not simply on g but on some other factor which Alexander (1935) called X . As this extra factor is supposed to involve 'industriousness and interest' it can, I think, be argued with some plausibility that a test such as the VRT is unlikely to be as good a test of academic potential as, say, a test such as CMT, even if it is a good test of g . This is because 'industriousness and interest' seems likely to reveal itself pre-eminently in a person's intellectual acquisitions, meaning by this not only the range of relatively specific skills which he acquires but also, of course, his vocabulary, general knowledge, and so on, things which, we have seen, the CMT taps, though not, apparently, very satisfactorily. The attractiveness of the VRT, on the other hand, is at least partly due to its relative independence of the effects of different kinds and degrees of education. Success on it appears to depend on the possession of one or two rather general skills which most children are not taught, or not taught formally, at school, namely, of course, the reasoning skills of drawing the appropriate inferences from a set of facts and of being able to tell whether a conclusion may legitimately be drawn from certain premises.

Ultimately, of course, a judgement as to the adequacy of the VRT, either as a means of selecting candidates for places at university or training college or as a measure of general intellectual ability, must rest on empirical evidence of the kind offered, in a modest form, in section 1.6. Since no test is perfect as an instrument for selection purposes, there will always be individuals who end up on the wrong side of the line dividing those who are selected from those who are rejected, so that it is in general no objection to a test to point to individuals whose performance on the relevant criterion (class of Honours, for example) is much better (or worse) than their performance on the test would lead one to expect. Exceptions to this rule, I think, would be cases where there are independent grounds for believing that the results of the test accurately reflect the subjects' standing on the abilities measured by the test, for this casts doubt on the alleged theoretical relationship between score on the test and performance on the criterion.

1 This is not, of course, an uncommon procedure in this area of psychological testing. On the contrary, tests of intelligence have so often been 'validated' in this way that it has sometimes been suggested (Anastasi, 1968) that they should simply be referred to as tests of scholastic aptitude.

The body of this thesis took as its focus the case of subjects whose performance on Section B of the two forms of the VRT - a section which, as we have seen, contributes two-thirds to the total score - was markedly lower than their academic attainment and ability measured in other ways, the whole trend of the results of the experiments described in that part being to confirm their relative weakness in the general area of reasoning. One of these subjects (I. McA. in Appendix E) whose score on Section B of the VRT was a little below the average of the Ordinary Psychology class as a whole has since graduated with first class honours in chemistry. Another PR subject (J.S.), whose Section B score was somewhat lower than I. McA.'s, graduated with a good second class honours degree in classics at the end of his Junior Honours year, instead of one year later, as he would be expected to do. Interestingly enough, Valentine himself may have come across a case of this kind, for in his Handbook to the test he refers to a graduate with first class honours in history who scored 9 on the VRT as a whole. Valentine (1954, p. 20, footnote) accounts for this case in terms of a 'general carelessness and indifference', citing, as evidence, the subject's failure to follow the instructions in item 2. Of course, he may be right about this, in which case the subject's score would not represent his true standing on the ability measured by the VRT and so would not count against the assumptions of that test; on the other hand, it is surely exceedingly rare for a person willingly to lose face if he could easily avoid doing so.

There are more general grounds for suspecting that a test which depends as heavily as Valentine's apparently does on something which is most naturally described as logical reasoning ability may be less than adequate as a means of assessing general intellectual ability or, insofar as the one depends on the other, academic potential. According to Guilford (1969) logical reasoning or, as he would prefer to call it, since in most tests with a heavy loading on this factor the subject is not asked to reason himself but to judge the validity of an argument presented, logical evaluation, was recognised as a special ability as long ago as 1938 when Thurstone published the first results of his factorial studies of human ability and has been repeatedly confirmed by factor analysts since that time. And whatever his doubts about some of the factors identified by Guilford and his colleagues Vernon (1950) seems to be prepared to accept logical reasoning as at least a minor group factor - to the extent of finding a place for it in his diagram of factors in psychological tests (ibid., p. 83).

Valentine himself expresses doubts about the effects of logic instruction on scores in Section B, going so far as to say that he thinks it 'probable that a training in symbolic logic or in the detection of syllogistic fallacies might have an appreciable effect on scores' in that section (1954, p.25). The fact that Elton's results (1965) seem to suggest that this is not the case (a sixteen week college course on elementary logic seemed to have no effect on scores in Section B) cannot, I think, be regarded as sufficient rebuttal of the view I have been taking: differences in logical reasoning ability of the kind required for success in the relevant section of the VRT do not seem to be produced by differences in formal education - although, as I have occasion to remark in a later chapter, the truth is that very little is known about the way in which sensitivity to the logical aspects of a thought product develop.

What I have been suggesting is that there are grounds for doubting the theoretical basis of the claim that performance on a test of inductive and deductive reasoning in which the latter element is predominant could be adequate as a test either of general intellectual ability or, therefore, of academic potential: in the latter context it seems highly likely that such a test will be unfair to some subjects, not simply because of the inherent limitations of any selection instrument but because the ability on which it depends is not evenly distributed over all subjects of comparable general ability. The limitations on the test's use which this view implies do not apply, of course, to its use as a means of assessing inductive and, especially, deductive reasoning. It is true that there are other tests of these abilities (the Critical Thinking Appraisal of Watson and Glaser (1964), for example, though in the light of Ennis's review (1962) it appears to have serious faults of its own) but none of them appear to test the same critical aspects of deductive reasoning ability as the VRT Section B. The traditional syllogism test, of which Hertzka and Guilford's Logical Reasoning test (1955) is a recent example, does not require the subject to identify the weakness in an argument and frequently this is the most difficult part of the task, just as it is the best evidence that the fallacy is understood. It is for its potential as a research instrument in this area that I regard the Valentine test as worthy of the effort involved in a revision and, of course, in the development of an alternative form.

A final word ought, perhaps, to be said about the implications of

the criticisms I have made of the VRT as it exists at present for the research described in the main part of this thesis. It might be said that a structure is only as strong as the foundations on which it rests and that as the division of subjects into 'poor reasoners' and 'good reasoners', which is fundamental to that research, was based on performance on the faulty Section B of the VRT, attempts to discover the factors which are responsible for this kind of difference could scarcely hope to meet with success. To this I think it is enough to reply, first, that the defects of the Valentine test became known to me only after the selection of subjects had been made and, indeed, the experimentation carried out - so that mine was not a decision to use an admittedly imperfect instrument for the purpose in hand; secondly, and perhaps more important, while the defects of the test might very reasonably have been held to explain a failure to find any differences of a relevant kind between the groups constituted by reference to performance on the VRT Section B, they are of relatively little importance when, as we have seen, the evidence of the experiments constantly confirms the existence of a difference of the kind assumed and when it is possible to establish meaningful relationships between this difference and other differences experimentally established (in particular between a relatively poor performance on the VRT Section B and proneness to the 'converse error'). It has not, of course, been the contention of this appendix that the VRT as a whole or Section B in particular is worthless either as a means of distinguishing those with high, from those with low, academic potential or as a way of separating those with good from those with 'poor' logical reasoning ability: it is simply that it could clearly be better. Perhaps if I had recognised its deficiencies before using it as a basis for selection, I might have obtained a clearer cut result by disregarding performances on items 7, 15A and 15B. To admit this, however, is obviously not to admit to any doubts as to the general validity of the VRT Section B as a measure of logical reasoning ability.

MATERIALS AND DATA RELATING TO THE VALENTINE REASONING TESTS

REASONING TESTS

FOR HIGHER LEVELS OF INTELLIGENCE

Prepared by C. W. VALENTINE, M.A., D.PHIL.
Emeritus Professor of Education, University of Birmingham

Name.....

Sex..... Age.....

Address

.....

FOR THE MARKER	
Score :—	
Section A
Section B
TOTAL	<u>.....</u>

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO

This booklet contains two sets of reasoning tests.
Section A has four tests, Section B has twelve.

Write your answers in the places provided.

Unless you are told otherwise, you will be allowed **55** minutes for the whole series—Section A and Section B.

Section A should be done first.

At the end of twenty minutes a warning will be given, and you should then proceed to Section B, even if you have not finished Section A. You can return to Section A later if you have finished Section B before the time is up. If you have done all you can of Section A before the warning given twenty minutes from the start, you should begin Section B at once without waiting.

The tests should be attempted in the order given, as in each series the harder tests come towards the end. But do not spend too long on any one test. Pass on to the next and return to the previous test later if you have time.

OLIVER AND BOYD LTD. EDINBURGH - - - TWEEDDALE COURT
LONDON - - - 39A WELBECK STREET, W.1

SECTION A

1. John Huggins was found shot dead in Church Street at 1.45 p.m. He had two known enemies, Bill Frogger and Jack Toper. The doctor says it could not have been suicide and that Huggins died between 12.30 p.m. and 1 p.m. the same day.

Frogger was seen running from Church Street by two reliable witnesses at 1.10 p.m.; Toper was seen by two reliable witnesses two miles from Church Street at 1.10 p.m.; and Frogger was seen by two other reliable witnesses in a public house $1\frac{1}{2}$ miles away from Church Street at 12.30 p.m. Neither Frogger nor Toper had any means of transport except their own legs. The maximum speed of running for each was 1 mile in 8 minutes.

Assuming that none but the two named could have committed the murder, underline what *must* be true in the following statements, cross out what *must* be false, and put a question mark by those which may or may not be true.

- Frogger (a) was the murderer
(b) could have been the murderer
(c) could not be the murderer .

- Toper (a) was the murderer
(b) could have been the murderer
(c) could not be the murderer .

2.

At a dinner Mr A had soup, fish and cheese
Mr B had soup and fish but no cheese
Mr C had fish and cheese, but no soup

Nothing else was eaten or drunk. Later Mr A and Mr C developed food poisoning. (We have no report yet about Mr B.)

Assuming that the cause of the poisoning was in the dinner mentioned, underline those of the following statements which are certainly true, cross out those which must be false, and put a question mark [?] by those which may or may not be true.

There was

- | | |
|---|--------------------------|
| (i) poison in the soup | <input type="checkbox"/> |
| (ii) poison in the cheese | <input type="checkbox"/> |
| (iii) no poison in the fish | <input type="checkbox"/> |
| (iv) poison in the soup and the fish | <input type="checkbox"/> |
| (v) no poison in the fish or the cheese | <input type="checkbox"/> |
| (vi) poison either in the fish or in the cheese | <input type="checkbox"/> |
| (vii) poison in the soup, the fish and the cheese | <input type="checkbox"/> |
| (viii) poison in the fish and the cheese | <input type="checkbox"/> |

3. There are four towns, A, B, C, and D. A is the same distance from B that B is from C, and C is half that distance from A.
D is the same distance from C that C is from A.
Is A nearer to C or to D, or the same distance from each ?

Diagrams may be used.

4. A man left his money to his five sons as follows : find the scheme or principles on which he divided the money, the individual personalities not being considered.

To A, aged 35, with 2 children, and income of £400, he left £500
To B, aged 40, with 3 children, and income of £500, he left £700
To C, aged 45, with 1 child, and income of £400, he left £600
To D, aged 35, with 2 children, and income of £600, he left £500
To E, aged 30, with no children, and income of £300, he left £200

(Half marks will be awarded for a partial solution of this problem, i.e. if only one principle or rule is discovered. For the full solution precise figures must be given.)

SECTION B

Examine the following arguments and state whether they are sound or not. You must *assume* first that the given premisses (*i.e.* the statements underlined) are *true*. The problem is, in each case, this : granted that these assumptions are true, is the other statement necessarily true? If you think the argument is sound, underline " Yes " and cross out " No " ; if the argument is unsound, cross out " Yes " and underline " No ".

If you say the argument is *unsound*, show which of the sentences, (i to iii or iv) given below that argument, gives the best reason why the conclusion does not follow from the given premisses. Mark your selected reason with a **X** in the blanks provided.

You are advised first to decide for yourself whether the argument is sound or not, before examining the reasons given below it. If the argument is *sound* it is obviously useless to examine the reasons which follow it.

Remember that you must assume that the underlined statements are true.

NOTE.—Marks will be deducted for wrong answers, so that mere guessing is penalized.

5. " All successful authors are industrious. John Smith is an industrious author. Therefore he is or will be a successful author."

- (i) It is not true that all successful authors are very industrious.
- (ii) The fact that all successful authors are industrious does not imply that all industrious authors are successful.
- (iii) Some successful authors are both industrious and clever.

Yes.	<u>No.</u> (i) (ii) (iii)
-----------------	---

6. “Everyone is either well informed of the facts or already convinced on the subject ; no one can be at the same time both already convinced on the subject and amenable to argument ; hence it follows that only those who are well informed of the facts can be amenable to argument.”

- (i) This conclusion is the converse of the true one, for those who are well informed of the facts will be sure of their ground and so will *not* be amenable to argument.
- (ii) A man may be convinced on the subject, yet if a good argument is ably put before him, he may alter his opinion.
- (iii) There is no reason why everyone should be either well informed of the facts or already convinced on the subject ; there may be people who have never heard of the subject at all.
- (iv) The first premiss is not clear, for some may be both well informed of the facts and already convinced on the subject, and according to the second premiss, these will not be amenable to argument.

<u>Yes.</u>	No. (i) (ii) (iii) (iv)
-------------	--

7. “None but Whigs vote for Mr B. All who vote for Mr B. are ten-pound householders. Therefore none but Whigs are ten-pound householders.”

- (i) Only Whigs vote for Mr B., yet all Whigs need not do this, so that there may be Whigs who are not ten-pound householders.
- (ii) All those who vote for Mr B. are both Whigs and ten-pound householders, yet there may be ten-pound householders who do not vote for him, and hence need not be Whigs.
- (iii) Even if none but ten-pound householders vote for Mr B., that is not to say that some of the ten-pound householders do not vote for his opponent, and hence are not Whigs.
- (iv) There may be voters who are not Whigs, yet who vote for Mr B. on personal rather than on political grounds, and these will also be ten-pound householders.

Yes.	<u>No.</u> (i) (ii) X (iii) (iv)
-----------------	---

8. "If you argue on a subject which you do not understand, you will prove yourself a fool ; for this is a mistake fools always make."

- (i) The statement is not sufficient to define "a fool". Only one characteristic is given, and a man may not be a fool only because he makes this mistake, but for other reasons too.
- (ii) This argument is unsound, because a wise man may be able to argue on a subject which he does not understand without giving himself away, while a fool could not.
- (iii) It is not logical to conclude that a man is a fool because he acts like one in this one particular instance.
- (iv) Although fools always make this mistake, it is not stated that all who make this mistake are fools, so that others who are not fools may do so too.

Yes.

No. (i) (ii) (iii) (iv) ~~.....X.....~~

9. "All the students are either industrious or intelligent. Either industry or intelligence will ensure success in the examination. So all the students will pass the examination."

- (i) This conclusion is incorrect, for a student may fail in the examination through misfortune ; for instance, he may feel unwell when the examination takes place.
- (ii) A student may be industrious or intelligent, yet may be disqualified in the examination for bad conduct, e.g. copying from other candidates.
- (iii) Either industry or intelligence alone is surely insufficient for a success. A combination of the two is needed.
- (iv) If a student is neither industrious nor very intelligent he may pass if by good fortune he is asked questions bearing on the little knowledge he has.

Yes.

~~No.~~ (i) (ii) (iii) (iv)

10. “ This pamphlet contains seditious doctrines. The spread of seditious doctrines is dangerous to the State. Therefore this pamphlet must be suppressed.”

- (i) It is not stated that everything dangerous to the State must be suppressed, so in the premisses given there is no reason for the suppression of the pamphlet ; and in any case it is not stated that the pamphlet would spread seditious doctrine.
- (ii) The spread of seditious doctrines is not always dangerous to the State, for if the State is stable seditious doctrines will not affect it.
- (iii) The conclusion is incorrect, for the doctrines in the pamphlet may only appear to be seditious in the opinion of some people. Others may not consider them so.
- (iv) To suppress the pamphlet may not of itself avert the danger. The doctrines expressed in it can still be spread verbally by their originators, so other measures may also be necessary.

Yes.	<u>No.</u> (i) X (ii) (iii) (iv)
-----------------	---

11. “ If all the accused were innocent, some at least would have been acquitted. We may infer then, that none were innocent, since none have been acquitted.”

- (i) The innocent are often condemned to suffer for the guilty. Condemnation is no proof of guilt.
- (ii) The guilt of some may have placed all in a bad light, so that none would be acquitted.
- (iii) If only some of the innocent would be acquitted in any case, then some were not acquitted when they ought to have been.
- (iv) We are only told that some of the accused would be acquitted if all were innocent. The number innocent may have been less than all, and so insufficient to secure any acquittals.

Yes.	<u>No.</u> (i) (ii) (iii) (iv) X
-----------------	---

12. “No schoolboy can be expected to understand Constitutional History, and none but schoolboys can be expected to remember dates ; so that no one can be expected both to remember dates and to understand Constitutional History.”

- (i) We cannot assume that this conclusion is true. College students may easily do both, for they are sufficiently developed intellectually to understand Constitutional History, while they are not old enough to have forgotten the dates they learned at school.
- (ii) One cannot say that no schoolboy can be expected to understand Constitutional History. A boy who is intelligent and well taught may easily do so.
- (iii) Schoolboys should not be expected to remember dates exactly, for this is an unsound method of teaching History, so that the whole argument is invalidated.
- (iv) The premisses are incomplete. No mention is made of schoolgirls, who are able to remember dates as well as boys.

Yes.

~~No.~~ (i) (ii) (iii) (iv)

13. “No soldiers should be brought into the field who are not well qualified to perform their part ; none but veterans are well qualified to perform their part ; therefore, none but veterans should be brought into the field.”

- (i) If only veterans were brought into the field, young soldiers would never have a chance to learn, and when the veterans died there would be no one to replace them. The conclusion is, therefore, unsound.
- (ii) Soldiers could not become veterans without going into the field as recruits, so the whole argument is false.
- (iii) It is a misstatement to say that none but veterans are well qualified to perform their part, since young soldiers make up for their lack of experience by their enthusiasm.
- (iv) Veterans may not be well qualified to perform their part for they may be too old, in which case the conclusion is invalid.

Yes.

~~No.~~ (i) (ii) (iii) (iv)

14.

“None but those who are contented with their lot in life can justly be considered happy. But the truly wise man will always make himself contented with his lot in life, and, therefore, it follows that he may justly be considered happy.”

- (i) A wise man can force himself to be contented with his lot in life, but the very fact of this compulsion will prevent his being truly happy.
- (ii) Those who are not content with their lot in life are often happy, for there is often more happiness in striving to attain one’s desire than in the actual attainment.
- (iii) The fact that only those who are contented with their lot in life can justly be considered happy, does not imply that all those who are contented with their lot are of necessity happy.
- (iv) The conclusion may be true or not ; it will depend on the standard of happiness. Content and happiness are not the same thing.

<u>Yes.</u>	<u>No.</u> (i) (ii) (iii) <u>X</u> (iv)
-------------	---

15.

“In Tutland only Conservatives—and not all of them—are Protectionists (i.e. against Free Trade) ; only Liberals—and not all of them—are Home Rulers : but both parties (Conservatives and Liberals) contain supporters of Women’s Franchise.”

It may be assumed that :

- 1. no Liberal is a Conservative ;
- 2. all who are not Protectionists are Free Traders ;
- 3. all who do not support Home Rule are Unionists.

This is all that is known about the views of the Tutlanders.

Hence with only this information

- (A) it would be *incorrect* to conclude that only the Unionists are Protectionists ;
- (B) it would be *wrong* to conclude that both Unionists and Free Traders are to be found among the supporters of Women’s Franchise.

(Note that there are *two* problems here. You have first to decide whether conclusion (A) is right or wrong ; and if wrong, which of the reasons under (A) opposite, best applies. Then you have to decide whether conclusion (B) is right or wrong ; and if wrong, which of the reasons under (B) opposite, best applies.)

- (A) (i) Unionists need not be Conservatives ; they may be Liberals, so that although Conservatives are Protectionists, Unionists need not be.
- (ii) In the given premisses, Unionists are only found in the Liberal Party which contains no Protectionists. Thus Unionists cannot be Protectionists.
- (iii) Only the Liberals are Home Rulers. Therefore all the Conservatives must be Unionists. As only Conservatives are Protectionists, it follows that only Unionists are Protectionists.
- (iv) The questions of Unionism and of Protection are entirely independent of each other. Thus no conclusion can be drawn as to whether believers in Protection are also believers in Unionism.

Yes : it <i>would</i> be incorrect.
<u>No</u> : it would <i>not</i> be incorrect. (i) (ii) (iii) X (iv)

- (B) (i) The Liberals who support Women's Franchise may be those who are Home Rulers, and the Conservatives just those who are Protectionists. Thus there may be neither Free Traders nor Unionists among the supporters of Women's Franchise.
- (ii) Since all Liberals are Free Traders and all Conservatives are Unionists, then there must be both Unionists and Free Traders among those who support Women's Franchise.
- (iii) The Liberals who support Women's Franchise may be all Free Traders, and the Conservatives who support Women's Franchise may be those who are not Protectionists, so that these supporters may all be Free Traders.
- (iv) The question of Women's Franchise is one which is not affected by considerations of Unionism or Free Trade, so these considerations are irrelevant.

Yes : it <i>would</i> be wrong.
<u>No</u> : it would <i>not</i> be wrong. (i) (ii) X (iii) (iv)

REASONING TESTS

for Higher Levels of Intelligence

FORM W

Test Booklet

Do not open this booklet until you are told to do so.

This booklet contains two sets of reasoning tests.
Section A has four tests, Section B has twelve.

Write your answers in the places provided in the Answer Booklet.

Unless you are told otherwise, you will be allowed 55 minutes
for the whole series - Section A and Section B.

SECTION A SHOULD BE DONE FIRST.

At the end of twenty minutes a warning will be given, and you
should proceed to Section B, even if you have not finished
Section A. You can return to Section A later if you have
finished Section B before time is up. If you have done all
you can of Section A before the warning given twenty minutes
from the start, you should begin Section B at once without
waiting.

The tests should be attempted in the order given, as in each
series the harder tests come towards the end. But do not
spend too long on any one test. Pass on to the next and
return to the previous test later if you have time.

Do not write anything in this test booklet.

SECTION A.

1. Smith and Jones both intended to put their names down for the annual Hill Race. This involved being present in person to sign their names on a sheet of paper posted on a notice board in the village hall at Cairncry (where the race began and finished) between 9 a.m. and 5 p.m. on the 13th April.

All we know about their movements is as follows: at 3 p.m. on the 13th April Smith was striding up and down impatiently at his sister's house in Dalgleish, 180 miles from Cairncry. At that very moment Jones was changing a wheel on his car at the roadside at a point 40 miles from Cairncry. Road transport is the fastest means of travel to and from Cairncry and, owing to the character of the roads in that region, the highest possible average speed between any two points within a 200-mile radius of Cairncry is 35 m.p.h.

Granted only that Smith and Jones had both intended to sign the sheet at Cairncry between 9 a.m. and 5 p.m. on the 13th April, under (1) in the answer booklet, underline those statements which must be true, cross out those which must be false and put a question-mark (?) against those which may or may not be true.

2. Suppose that a car will fail to start if and only if there is no petrol in the tank or if the petrol pump is broken or if the starter motor is jammed.

Mr. Z's car will not start although he has just filled the tank with petrol. Under (2) in the answer booklet underline those statements which, on the above supposition, Mr. Z can safely assume to be true, cross out those he can assume to be false and put a question-mark against those about which he cannot be sure.

3. A, B, C and D are four towns. D is the same distance from A as from B. The distance between A and C is likewise the same as the distance between B and C. And the distance between A and C is only half the distance between A and B. Is D the same distance from A and B as C is or farther away or nearer?

Diagrams may be used. These, together with your answer, should be written under (3) in the answer booklet.

4. A man has five children with varying educational records. By an outstanding coincidence they are all taking up their first posts at the same time, but with rather different starting salaries. Determine those principles the operation of which accounts for the differences in starting salary, given that all the relevant factors are included in the information about the man's children.

Ann attended a secondary school for 3 years, and had no further full-time education. Her starting salary is £450.

Bruce attended a secondary school for 6 years and a University for 4 years. His starting salary is £750.

Colin attended a secondary school for 5 years and had no further full-time education. His starting salary is £450.

Dora attended a secondary school for 4 years and a Commercial College for 3 years. Her starting salary is £600.

Emma attended a secondary school for 5 years and a University for 5 years. Her starting salary is £700.

(Half marks will be awarded for a partial solution of this problem, i.e., if only one principle or rule is discovered. For a full solution precise figures must be given).

Write your answer under (4) in the answer booklet.

200

SECTION B.

Examine the following arguments and state whether they are sound or not. You must assume first that the given premises (i.e. the statements underlined) are true. The problem is, in each case, this: granted that these statements are true, is the other statement (the conclusion of the argument) necessarily true? If you think the argument is sound, underline "Yes" in the answer booklet and cross out "No"; if the argument is unsound, cross out "Yes" and underline "No".

If you say the argument is unsound, show which of the sentences, (i to iii or iv) given below that argument, gives the best reason why the conclusion does not follow from the given premisses. Mark your selected reason with a X in the blanks provided in the answer booklet.

You are advised first to decide for yourself whether the argument is sound or not, before examining the reasons given below it. If the argument is sound it is obviously useless to examine the reasons which follow it.

REMEMBER THAT YOU MUST ASSUME THAT THE UNDERLINED STATEMENTS ARE TRUE.

NOTE:- Marks will be deducted for wrong answers, so that mere guessing is penalized.

5. "Every wise leader is a good listener. Mr. Jones is a good listener. Therefore he is or will be a wise leader".
- (i) Some wise leaders are both good listeners and firm of purpose.
 - (ii) It is not true that every wise leader is a good listener.
 - (iii) The fact that every wise leader is a good listener does not imply that every good listener is a wise leader.
6. "Only people who are interested in major sporting events are admirers of Mr. C. All Mr. C's admirers are people who enjoy strenuous physical exercise. Hence only people who are interested in major sporting events are people who enjoy strenuous physical exercise".
- (i) Even if it is only people who enjoy strenuous physical exercise who admire Mr. C. that is not to say that some people who enjoy such exercise are not among his supporters and therefore not interested in major sporting events.
 - (ii) Only people interested in major sporting events are admirers of Mr. C, yet all people with this interest need not be his admirers, so there may be people interested in major sporting events who do not enjoy strenuous physical exercise.

- (iii) All Mr. C's admirers are interested in major sporting events and also enjoy strenuous physical exercise, yet there may be people who enjoy such exercise but who are not admirers of Mr. C. and who therefore need not be interested in major sporting events.
- (iv) There may be people who are not interested in major sporting events and who nonetheless admire Mr. C for reasons unconnected with sport, and these will also be people who enjoy strenuous physical exercise.

7. "Everyone is either already engaged in active military service or else convinced that we shall shortly win this war. No one can be at the same time convinced that we shall shortly win this war and in need of some boost to his morale. Therefore only those already engaged in active military service are in need of a boost to their morale".

- (i) The first premiss is not clear, for some may be both already engaged in active military service and convinced that we shall shortly win this war, and according to the second premiss these will not require a boost to their morale.
- (ii) This conclusion is the reverse of the correct one, for those who are already engaged in active military service will know how the war is going and so will not need to have their morale boosted.
- (iii) It may be true that everyone is either already engaged in active military service or convinced that we shall shortly win this war: there is always in any country a minority who are pacifists or who doubt the competence of those responsible for the conduct of the war.
- (iv) A person may be convinced that we shall shortly win the war and still need, in moments of depression, a boost to his morale.

8. "This candidate is a pacifist. Anyone who advances the cause of peace is contributing to international understanding. Therefore this candidate should be elected".

- (i) Electing the candidate may not in fact be a contribution to international understanding. His ability to persuade people of the desirability of peace may be very limited, so that some other means to this end may be necessary as well.
- (ii) It is not stated that everyone who contributes to international understanding should be elected, so there is nothing in the premisses which justifies the election of the candidate in question; and in any case it is not stated that the candidate does advance the cause of peace.
- (iii) The conclusion is incorrect, for the candidate may appear to be a pacifist only in our typically aggressive society - in other societies he might even be thought a warmonger.
- (iv) Advancing the cause of peace is not always contributing to international understanding for there are occasions on which the former can be achieved only at the expense of the latter.

9. "All the prisoners are either emotionally unstable or mentally defective. Either emotional instability or mental defect makes a person ill-equipped to cope with the stresses of modern life. So all the prisoners are ill-equipped to cope with the stresses of modern life".

- (i) A prisoner may be emotionally unstable or mentally defective and still manage to cope, given the constant support of some other person - for example his wife.
- (ii) This conclusion is incorrect, for a prisoner may be adequately equipped to deal with the stresses of modern life in some other way - he may draw compensating strength from his religion, for example.
- (iii) Even if a prisoner is neither emotionally unstable nor mentally defective he may still not be adequately equipped for modern life if he is born into a particularly difficult segment of it.
- (iv) Surely neither emotional instability nor mental defect by itself is enough to unfit a man to cope with modern life; a combination of the two is required.

10. "If you drive a car without first ensuring that it is in satisfactory mechanical order, you will reveal your inexperience as a driver for inexperienced drivers are always guilty of this oversight".

- (i) Although inexperienced drivers always fail to ensure that their car is in satisfactory mechanical order before driving it, it is not stated in the argument that all who are guilty of this oversight are inexperienced drivers, so that others who do not lack driving experience may be guilty of this oversight too.
- (ii) No satisfactory explanation of what is meant by 'inexperienced' is given. A person may be judged to lack experience as a driver not only because he is guilty of the kind of oversight mentioned but because of other things he does as well.
- (iii) It is not logical to conclude that a person lacks experience as a driver because he behaves in one respect as if he does.
- (iv) This argument is invalid, because an experienced driver may be able to cope with the consequences of mechanical defects in his car when an inexperienced one would not.

11. "If everyone in this room had felt it too warm, surely someone would have complained. Since no one did complain, we may conclude that no one felt it too warm".

- (i) Perhaps everyone felt that if the others could suffer in silence he could too and this is why no one complained.
- (ii) If only one person would have complained in any case, then some did not complain when they had reason to do so.
- (iii) People in a group are often too self-conscious to complain. The fact that a person does not complain does not show that he is not uncomfortable.
- (iv) We are told only that someone would have complained if everyone had felt it too warm. Perhaps not everyone felt it too warm, and that is why no one complained.

12. "No teacher should be put in charge of a difficult class unless he can be relied upon to keep it under control. Only male teachers with some experience of such classes come into this category. Therefore only such teachers should be put in charge of a difficult class".

- (i). All teachers who have had some experience of difficult classes must at one time have been teachers without such experience, so the whole argument is false.
- (ii) If only male teachers with some experience of difficult classes were put in charge of such classes, other male teachers would never have a chance to gain experience of these classes, and when their seniors retired there would be no one to replace them. The conclusion is, therefore, unsound.
- (iii) Teachers with experience of difficult classes may not be able to keep them under control, for they may go into the classroom overwhelmed by their knowledge of the difficulties that await them. In that case the conclusion is invalid.
- (iv) It is a mistake to think that only teachers with some experience of difficult classes can be relied upon to keep them under control, for inexperienced teachers possess the resilience of youth and so are better fitted to respond to the demands of a difficult classroom situation.

13. "Only those who are free from any taint of self-interest can be entrusted with the role of judge. Those who take to heart the teachings of the great religious and spiritual leaders of mankind lose all trace of self-interest. Hence such people can be entrusted with the role of judge."

- (i) The conclusion may or may not be true - it will depend on what we expect from a judge. Justice is not the same thing as disinterest.
- (ii) A person who is preoccupied with religious and spiritual doctrines is frequently too little interested in everyday affairs to make a good judge of these things.
- (iii) Those who are occasionally subject to a temptation to further their own interests at the expense of others will understand the motives of the men they are called upon to judge better than those who are never subject to such temptations.
- (iv) The fact that only those who are free from any hint of self-interest can be entrusted with the role of judge does not imply that all those who have this quality can be entrusted with the role in question.

14.

251

"No gourmet can be expected to enjoy Cottar's Curd and only a gourmet can be expected to enjoy Camembert. So no one can be expected to enjoy both Cottar's Curd and Camembert".

- (i) The premisses are not complete. No mention is made of professional buyers in the dairy-produce industry who enjoy a good Camembert as much as gourmets do.
- (ii) It is a mistake to say that no gourmet can be expected to enjoy Cottar's Curd. A genuine gourmet recognises that even the humbler foods have a place in the whole range of cookery.
- (iii) We cannot assume that this conclusion is true. There are people who enjoy Cottar's Curd because of the sentimental associations it has and on the other hand enjoy Camembert simply for its taste.
- (iv) The gourmet's liking for Camembert is not genuine but simply a reflection of his responsiveness to prevailing fashions - so that the whole argument is invalidated.

15.

"In Abasiland only the elderly - and not all of them - are wise and only the young - and not all of them - are healthy. However, both the elderly and the young include some individuals who are happy.

It may be assumed that as far as Abasiland is concerned:

- (1) all who are not wise are lacking experience of the world.
- (2) all who are not healthy are physically weak.

On the basis of this information alone

- (A) it would be incorrect to conclude that in Abasiland only those who are physically weak are wise;
- (B) it would be wrong to conclude that in Abasiland both the physically weak and those lacking in experience of the world include some individuals who are happy.

(Note that there are two problems here. You have first to decide whether conclusion (A) is right or wrong; and if wrong, which of the reasons under (A) below best applies. Then you have to decide whether conclusion (B) is right or wrong; and if wrong, which of the reasons under (B) below best applies).

- (A) (i) In the given premisses the physically weak are found only among the young, none of whom are wise. Thus the physically weak cannot be wise.
- (ii) Only the young are healthy. Therefore, all the elderly must be physically weak. As only the elderly are wise, it follows that only the physically weak are wise.
- (iii) The question whether a person is physically weak is entirely separate from the question whether he is wise. Thus no conclusion can be drawn as to whether the wise are also physically weak.
- (iv) The physically weak need not be elderly; they may be young, so that although the elderly are wise, the physically weak need not be.

- (B) (i) Since all young people in Abasiland are lacking experience of the world and all elderly people are physically weak, there must be both the physically weak and persons lacking experience of the world amongst those who are happy.
- (ii) The question of happiness is not one which is affected by a person's being physically weak or lacking in experience of the world, so these considerations are irrelevant.
- (iii) The young who are happy may be these who are healthy, and the elderly who are happy may be these who are wise. Thus there may be neither persons lacking in experience in the world nor persons who are physically weak amongst those who are happy.
- (iv) The young who are happy may all be lacking in experience and the elderly who are happy may be those who are not wise, so that those who are happy may all be lacking in experience.

REASONING TESTS

for Higher Levels of Intelligence

FORM W

Answer Booklet

Name

Sex

Class

Date of Testing

Do not open the Test Booklet until you are told to do so.

Please note the identification number in the top right hand corner of this booklet. You will be able to identify your score on the test, when it is given, only by reference to this number.

SECTION A.

1. Smith (a) signed his name before Jones.
(b) could have signed his name before Jones.
(c) could not have signed his name before Jones.

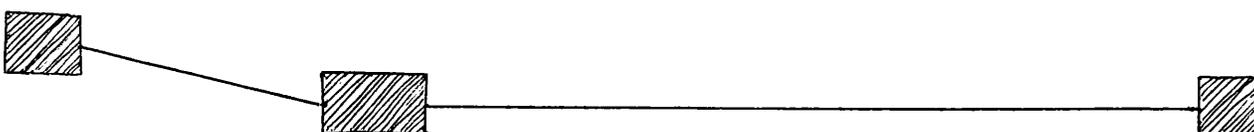
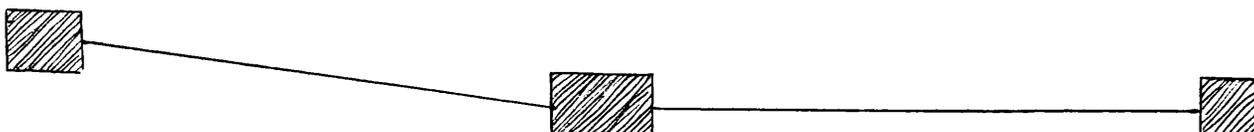
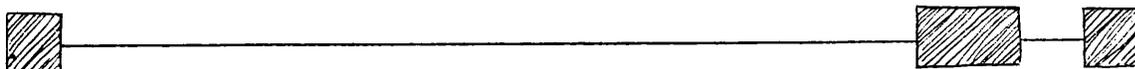
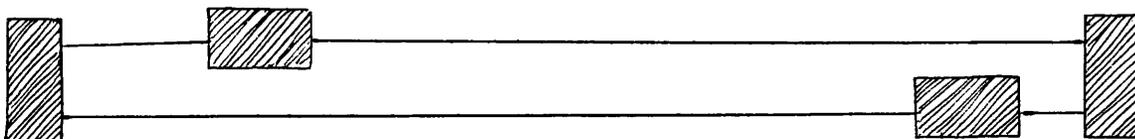
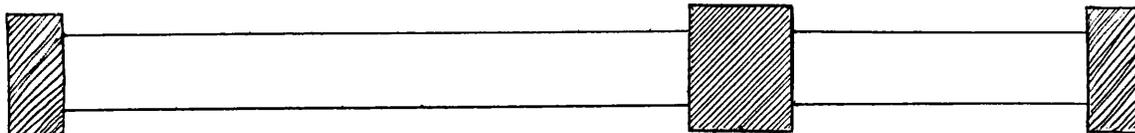
Jones (a) signed his name before Smith.
(b) could have signed his name before Smith.
(c) could not have signed his name before Smith.
2. (i) the petrol pump is broken.
(ii) the starter motor is jammed.
(iii) it can't be that there is no petrol in the tank.
(iv) the petrol pump is broken and the starter motor is jammed.
(v) either the petrol pump is broken or the starter motor is jammed.
(vi) if the starter motor is jammed then the petrol pump is not broken.
(vii) if the petrol pump is not broken then the starter motor is jammed.
(viii) neither the petrol pump nor the starter motor is responsible for the failure.
- 3.
- 4.

SECTION B.

5. ~~Yes~~ No (i) (ii) (iii) ~~..X..~~
6. ~~Yes~~ No (i) (ii) (iii) ~~..X..~~ (iv)
7. Yes ~~No~~ (i) (ii) (iii) (iv)
8. ~~Yes~~ No (i) (ii) ~~..X..~~ (iii) (iv)
9. Yes ~~No~~ (i) (ii) (iii) (iv)
10. ~~Yes~~ No (i) ~~..X..~~ (ii) (iii) (iv)
11. ~~Yes~~ No (i) (ii) (iii) (iv) ~~..X..~~
12. Yes ~~No~~ (i) (ii) (iii) (iv)
13. ~~Yes~~ No (i) (ii) (iii) (iv) ~~..X..~~
14. Yes ~~No~~ (i) (ii) (iii) (iv)
- 15A. ~~Yes~~: it would be incorrect
No: it would not be incorrect
(i) (ii) ~~..X..~~ (iii) (iv)
- 15B. ~~Yes~~: it would be wrong
No: it would not be wrong
(i) ~~..X..~~ (ii) (iii) (iv)

(d) VALENTINE REASONING TESTS, SECTION B (FORM W) SCORING STENCIL 1

Ensure that 'Section B' is visible
in this window: 



A 'pair' is an underlined 'No' with a cross in the window to the right along the line. For each pair put a tick in the window at the right hand end of the line.

Number of pairs:  x 5 = 

A 'single' is an underlined 'No' without an accompanying cross.

Number of singles: 

VALENTINE REASONING TESTS, SECTION B (FORM W) SCORING STENCIL 2

Ensure that 'Section B' is visible
in this window: 

Four horizontal lines, each with a hatched square at the left end and a hatched square at the right end, representing a window for marking.

Wherever the 'Yes' in the window is underlined, place a tick in the other window at the end of the line.

Number of Yeses underlined:



x 2 =

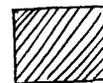


WRONGS

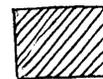
Ensure that 'Section B' is visible
in this window: 



Number of Yeses underlined:



Number of Noes underlined:



Total:



(e) Distributions of scores on the two forms of the Valentine test (M-Y and A-L groups separately) and on the Concept Mastery Test and the Scottish Certificate of Education (Arts and Science students separately).

<u>Valentine, Form V</u>			<u>Valentine, Form W</u>		
	<u>M-Y</u>	<u>A-L</u>		<u>A-L</u>	<u>M-Y</u>
70-71	1	1	70-71		1
65-69	2	2	65-69	1	5
60-64	5	10	60-64	4	13
55-59	8	15	55-59	12	17
50-54	11	23	50-54	17	27
45-49	15	24	45-49	22	23
40-44	32	24	40-44	25	23
35-39	17	20	35-39	21	17
30-34	10	24	30-34	23	12
25-29	19	7	25-29	17	14
20-24	17	6	20-24	10	3
15-19	13	4	15-19	6	3
10-14	6		10-14	1	1
5-9	2		5-9	2	
	158	160		161	159

<u>Concept Mastery Test</u>			<u>S.C.E. 'scores'</u>		
	<u>Arts</u>	<u>Science</u>		<u>Arts</u>	<u>Science</u>
170-9	2		20		1
160-9	3	1	19		0
150-9	7	2	18		1
140-9	21	2	17		0
130-9	14	5	16		2
120-9	25	12	15	3	4
110-9	28	11	14	3	0
100-9	31	14	13	4	4
90-9	36	17	12	10	7
80-9	19	10	11	7	2
70-9	18	15	10	18	7
60-9	8	5	9	27	7
50-9	3	4	8	20	13
40-9	1	4	7	35	12
30-9	2	2	6	20	6
	218	104	5	18	14
			4	12	3
			3	11	6
			2	10	5
<u>S.C.E.</u>	<u>Mean</u>	<u>S.D.</u>	1	1	
<u>Arts</u>	7.20	3.18	0	6	
<u>Science</u>	8.03	3.87		205	94

Notes As explained in the text, (1) the M-Y group did the V form of the VRT first, the A-L group the W form; (2) S.C.E. 'scores' were obtained by awarding 3 points for an A pass, 2 for a B and 1 for a C, subjects with a score of 0 having had no passes on the Higher grade in their fifth year at secondary school.

(f) Special instructions sheet used in the second VRT testing session, reproduced here because of its possible bearing on the extent to which subjects were likely to base their answers to the second Form which they attempted on their memories of their answers to the first. The instructions, in the form of a letter, were on the subjects' desks on their arrival in the examination hall.

Department of Psychology,
10th November 1967.

Dear Class Member,

As you will see, the third test in the series looks very much like the first - except for the colour of the paper it is printed on! In fact it is what psychologists call an alternative form of the same test - i.e., it is supposed to be a test of the same difficulty and of the same kind of difficulty as the first. The aim is to have a test which is as similar as possible to the first without being so like it that you can do the second by remembering your answers to the first. In fact in nearly every case an answer which was right in the first test you sat will be wrong in the second. So this is as much a test of intelligence and perseverance as the first was. Any test or examination can only sample the knowledge or ability it is intended to measure, and the two forms of the test, with an interval of a week between, ought to be a much better measure of the qualities concerned than a single form could be.

The questionnaire which you will find on your desk is intended to give you an opportunity to voice your opinions on the tests you have sat, and also to provide us with information about your attitude towards the tests and your way of tackling them. This information will be of great value in our attempts to understand what these tests measure and also how they are likely to be viewed by the people who sit them. If you would like to use the blank reverse sides of the questionnaire itself to make additional comments, I should be glad if you would do so.

If you have arrived early you may be able to answer all but the last 7 or 8 questions before you begin today's test. Equally, if you finish the test early you may be able to complete the questionnaire before the 55 minutes are up. If you can manage, would you please do this and leave the completed questionnaire, together with the test and answer booklets, on your desk when you leave the hall. If you do not have time to complete the questionnaire today, would you please bring it, completed, to the next meeting of the class - on Monday - when I shall arrange to have them collected.

Finally, I would like to say how impressed I have been by the punctuality and good nature of the class as a whole. I hope you have found the experience of sitting these tests interesting and enjoyable, and that you will say so - in the answers you give to the last question in the questionnaire - if you did not!

Yours sincerely,

Ian G. Wallace

Note It is perhaps as well to say that very few indeed of the class did attempt to complete the questionnaire in time which they might otherwise have spent on the test.

(g) The questionnaire. The percentages given down the right-hand side refer the numbers who actually replied to that question. 202 Arts and 98 Science students returned questionnaires. The average number of omissions for the two groups was just over 2 and just under 2. These were spread fairly evenly over all items with the following exceptions: for Arts students, items 12, 27 and 28 produced 8, 8 and 19 omissions; for Science students, item 28 also produced an unusually large number of omission, viz., 10. It will be apparent that the Science percentages generally approximate fairly closely to the actual numbers of students giving that reply and that a similar approximation for Arts students may be obtained by doubling the percentage figure. For the terms in which the questionnaire was introduced see the instructions sheet reproduced in (f) above.

VALENTINE'S REASONING TESTS AND TERMAN'S

CONCEPT MASTERY TEST: A QUESTIONNAIRE

Put a tick in the appropriate box:

		<u>Arts</u>	<u>Science</u>	
		<u>%</u>	<u>%</u>	
1	Name			
2	Sex	Male	28	70
		Female	72	30
3	Age	17 years	4	2
		18 "	17	18
		19 "	36	23
		20 "	32	38
		21 "	5	13
		21+ "	10	6
4	Is your native language English	Yes		
		No		
		Bilingual		
5	Faculty	Arts	67%	
		Science	33%	
6	Which year at University is this for you?	1st	17	4
		2nd	43	33
		3rd	39	41
		4th	1	20
		5th or more	-	2
7	Name of your Adviser of Studies			
8	Do you think the Concept Mastery Test (CMT) is a good test of all-round intellectual ability?	Yes	35	30
		No	51	60
		Don't know	14	10
9	If not, why not?			

		<u>Arts</u>	<u>Science</u>
		<u>%</u>	<u>%</u>
10	Do you think Valentine's Reasoning Tests (VRT) are a good test of all-round intellectual ability?	Yes 39 No 43 Don't know 18	49 35 16
11	If not, what is wrong with them in your opinion?		
12	Do you think CMT and VRT <u>together</u> constitute a good test of all-round intellectual ability?	Yes 60 No 18 Don't know 22	48 33 19
13	If not, why not?		
14	Do you hope to take an Honours degree?	Yes 31 No 59 Don't know 9	36 56 8
15	If so, in what subject?		
		<u>Answering Yes to 14</u>	
16	Do you think the CMT tests an ability which is essential or desirable for success in the Honours subject, if any, you hope to take?	Essential 15 Desirable but not essential 68 Neither 12 Don't know 5	14 49 34 3
		<u>Answering Yes to 14</u>	
17	Do you think the VRT tests an ability which is essential or desirable for success in the Honours subject, of any, you hope to take?	Essential 28 Desirable but not essential 57 Neither 7 Don't know 8	50 40 6 3
18	Are you good at spotting weaknesses in another person's arguments?	Very good 3 Quite good 59 Not very good 36 Hopeless 2 Don't know 2	2 69 23 2 3
19	Do you think some people are much better than you at spotting such weaknesses?	Yes 89 No 4 Don't know 7	86 3 11
20	Have you attended any classes in logic?	Yes, at school - Yes, at University 26 No 74	1 8 91
21	Have you recently read any books on logic in your spare time?	Yes 11 No 88 Not sure 1	6 91 3
22	If so, can you remember the author(s) and/or titles?		

Arts Science
% %

23	Are you taking the Ordinary Logic (General Philosophy) or Ordinary Moral Philosophy class this year?	Logic Moral Philosophy Neither		
24	Have you taken the Ordinary Logic or Ordinary Moral Philosophy class in a previous year?	Logic Moral Philosophy Neither		
25	Are you currently taking, or have you previously taken, the Higher class in either of these subjects?	Logic Moral Philosophy Neither		
26	If you are taking, or have taken, any of these philosophy classes, do you think you were helped by them in your answers to VRT?	Yes	20	2
		No	29	1
		Don't know	6	-
		Doesn't apply	45	97
27	If you have studied logic, in your spare time or in a class at school or University, were you able to use what you learned in the course of that study to help you with Section B of VRT?	Yes	14	5
		No	9	5
		Don't know	3	1
		Doesn't apply	74	88
28	In particular did you use any of the following:			
	(a) Substitution of symbols for terms		1	3
	(b) Euler's circles or Venn diagrams		1	10
	(c) Translation of sentences with 'only' into sentences without 'only'		5	-
	(d) Conversion, obversion, etc., of propositions		1	-
	(e) None of these		13	5
	(f) Doesn't apply (haven't studied logic)		79	82
29	Do you think you could have done better on VRT if you had had more time?	Yes	51	32
		No	42	57
		Don't know	8	10
30	If so, how much longer would you have liked to have?			
31	Would this have been spent on:	Section A	13	5
		Section B	12	7
		Both	31	22
		Doesn't apply	44	66
32	Did the mere existence of a time limit interfere with your ability to concentrate on the task?	Yes	48	44
		No	52	56
33	Did you find it easy to understand what you were supposed to do in VRT Section B?	Very easy	30	39
		Quite easy	60	53
		Not easy	10	8

		Arts	Science
		%	%
34	Although you are warned in the instructions to VRT Section B that guessing will be penalised, did you find yourself <u>obliged</u> to guess if you weren't to leave a blank?	Often 3 Occasionally 53 Never 45	3 41 55
35	In the instructions to VRT Section B it advises you to look at the reasons given below the argument only <u>after</u> you have decided that the conclusion follows from the premises. Did you in fact find it helpful to look at these reasons <u>before</u> you decided one way or the other?	Always 2 Often 8 Occasionally 69 Never 21	2 12 57 29
36	Did you find the CMT more or less enjoyable to do than the VRT?	More enjoyable 60 Less enjoyable 22 About the same 17	28 51 21
37	Do you think it would be fair to use the tests you sat as the <u>sole</u> criterion for admission to the <u>Arts</u> faculty of a University?	CMT alone: fair 5 unfair 90 Don't know 5 VRT alone: fair 3 unfair 94 Don't know 4 CMT and VRT together: fair 13 unfair 75 Don't know 12	6 76 18 4 76 20 16 62 22
38	Do you think it would be fair to use the tests you sat as the <u>sole</u> criterion for admission to the <u>Science</u> faculty of a University?	CMT alone: fair - unfair 82 Don't know 18 VRT alone: fair 5 unfair 73 Don't know 22 CMT and VRT together: fair 5 unfair 72 Don't know 23	- 96 4 7 88 5 6 85 9
39	Would it improve the present system if results on tests of this sort were taken into consideration in the selection of candidates for admission to the two Faculties <u>as well as</u> passes in the S.C.E., G.C.E. etc.?	<u>Arts:</u> Yes 68 No 19 Don't know 13 <u>Science:</u> Yes 45 No 21 Don't know 34	58 18 24 62 29 9
40	Do you think it would be fair to use these tests as the <u>sole</u> criterion in the selection of candidates for admission to an Honours course?	CMT alone: Fair 3 Unfair 87 Depends on Hons. course 7 Don't know 5	- 80 17 3

		Arts	Science	
		%	%	
40 (contd)	VRT alone: Fair	1	-	
	Unfair	86	72	
	Depends on Hons. course	9	23	
	Don't know	4	5	
CMT and VRT together:	Fair	2	3	
	Unfair	76	65	
	Depends on Hons. course	16	24	
	Don't know	6	8	
41 In the selection of applicants for admission to an Honours course, do you think results on such tests ought to be taken into account <u>in addition to marks</u> in the relevant class below Honours level?	Yes	51	49	
	No	32	36	
	Don't know	17	14	
42 Which form of the VRT did you do first?	Form V (white booklet)	52	52	
	Form W (pink and blue booklets)	48	48	
43 (a) If you did <u>Form V</u> first which Form did you find <u>easier</u> ?	Form V	6	14	
	Form W	52	41	
	About the same	44	43	
	(b) If you did Form W first which Form did you find <u>easier</u> ?	Form V	35	47
		Form W	14	20
		About the same	51	33
44 If you did find one easier can you suggest why this was so?				
45 Did you find one of the Forms <u>more interesting</u> than the other?	Form V	8	11	
	Form W	9	7	
	About the same	83	82	
46 If so, can you suggest why?				
47 Do you think that your performance on any of the tests was adversely affected by purely temporary circumstances? Specify the tes involved.	(a) below par physically (e.g., heavy cold)			
	(b) unable to concentrate because of emotional upset			
	(c) customary examination nerves			
	(d) other (specify if possible)			
	(e) none			
48 Did you feel that you did one of the tests <u>better</u> than the others?	Yes, CMT			
	Yes, VRT Form V			
	Yes, VRT Form W			
	No			

	Arts	Science
	%	%

49 Did you feel you did one of the tests . Yes, CMT
worse than the others? Yes, VRT Form V
Yes, VRT Form W
No

50 You will be told your scores on the three tests as
soon as marking is completed, and there will be an
opportunity to discuss the tests and the answers
given to this questionnaire in class and in private.
Taking this into account, would you say that the
time you have spent on this project was:

(a) profitably spent? Yes
No
Don't know

(b) enjoyably spent? Yes
No
Don't know

Note The results given in this Appendix are those which appear to have
the most direct bearing on the issues arising in Appendix A. Other
results, with reference to the GR and PR groups are presented in Appendix E.

(h) Scores and ranks on the VRT, Form V, CMT and Ordinary Class exams of the 15 Honours Psychology students in the original class of 320.

<u>Sub-</u> <u>jects</u>	<u>Honours</u>		<u>Valentine</u>			<u>Concept Mastery</u>			<u>Class Exam</u>	
	<u>'Score'</u>	<u>Rank</u>	<u>Score</u>	<u>z</u>	<u>Rank</u>	<u>Score</u>	<u>z</u>	<u>Rank</u>	<u>Score</u>	<u>Rank</u>
MB	35	15	41	-0.19	12½	107	-0.04	10	137	5½
MC	45	9	41	-0.19	12½	97	-0.41	14	132	9
RC	49	8	48	+0.40	10	132	+0.47	6	114	14
RG	83	1	43	-0.02	11	171	+2.35	1	141	4
CG	41	10	64	+1.78	2	109	+0.03	9	137	5½
RMcC*	61	6	43	+0.48	9	115	+0.64	5	154	2
AMcM	73	2	31	-0.38	14	140	+1.19	4	131	10
DM*	62	5	52	+0.75	5	87	-0.31	12	128	11½
JP	34	16	44	+0.55	7½	119	+0.40	7	117	13
DS	68	4	64	+1.99	1	157	+1.83	2	155	1
MS	50	7	51	+1.05	3	144	+1.34	3	145	3
CS	40	11½	50	+0.98	4	116	+0.29	8	135	7½
RS*	38	14	44	+0.55	7½	85	-0.38	13	108	15
TS*	72	3	46	+0.69	6	94	-0.07	11	128	11½
JT	40	11½	26	-0.74	15	91	-0.64	15	135	7½

* Science students

Note The ranks were based on z scores in the cases of the VRT and CMT to take account of two distorting factors: (1) the fact that A-L subjects (including DM) did the VRT Form V after exposure to Form W and (2) the fact that the CMT clearly favoured Arts students. The z scores are therefore based on the means and standard deviations of the M-Y and A-L groups in the original Ordinary class in the case of the VRT and of the Arts and Science groups in the case of the CMT. The Honours 'scores' were found by awarding the following numbers of points for answers (N=20) agreed by three examiners to be of the relevant class: 1st class, 5; upper 2nd, 3; lower 2nd, 2; 3rd class, 1; fail, 0. The class exam scores are the sum of percentage scores on two class exams at the Ordinary level.

(i) a. Types of statement in the VRT Section B. In some cases, which I have marked with an asterisk, the grammatical form of the statement is far removed from what I give as its logical form. In the case of item 10 it seems that any attempt at a translation would likely mislead rather than illuminate the situation. I give the item numbers and the statements in the order in which they appear in each.

- | | | | |
|----|--|-----|--|
| 5 | All X's are Y's
A is a Y
A is an X | 12 | No X's are Y's
Only X's are Z's
No one is both an X and a Z |
| 6 | Everyone is either an X or a Y
No one is both a Y and a Z
Only X's are Z's | 13 | No X's are Y's
Only Z's are Y's
Only Z's are Y's |
| 7 | Only X's are Y's
All Y's are Z's
Only X's are Z's. | 14 | Only X's are Y's
All Z's are X's*
All Z's are Y's* |
| 8 | If p then q
All X's are Y's | 15A | Only X's (but not all X's)
B are Y's
Only Z's (but not all Z's)
are A's
Some X's are B's*
Some Z's are B's*
No Z's are X's
All non-Y's are C's
All non-A's are D's
Only D's are Y's
Some D's are B's* and
some C's are B's* |
| 9 | All A's are either X's or Y's
All A's which are X's or Y's
are Z's*
All A's are Z's | | |
| 10 | Unclassifiable | | |
| 11 | If all X's are Y's then some
X's at least are Z's
No X's are Z's
No X's are Y's | | |

In summary: The universal affirmative occurs, in one form or another 12 times. This includes two 'universal-disjunctives'. The 'Only X's are Y's' type of statement occurs 10 times. The universal negative occurs 7 times. There are two, very different hypotheticals. The particular affirmative occurs five times, though only once in a straight forward form.

b. The exact form in which the conclusions of 13 items in the two forms of the VRT were presented to students who were asked to say whether they were true or false. The sheet used, with the instructions, is reproduced overleaf.

(i) b.(contd.)

The following are statements appearing as the conclusions of arguments in Valentine's Reasoning Tests. Because an important factor in determining whether a person reasons correctly or not is the truth or falsity of the conclusions of the arguments he is considering, it is very important to know whether the following statements appear to most students (for whom Valentine's test is intended) to be true or not.

Look at each of the statements carefully and try to decide whether it is true or false. In some cases you may not be able to make up your mind one way or the other. In such cases just write a question-mark in the left-hand margin. If you think the statement is definitely true or definitely false, write T or F respectively. If you think it is probably true or probably false, write T? or F? respectively.

- 1 Only those who are well informed of the facts are amenable to argument.
- 2 Only Whigs are ten-pound householders.
- 3 Arguing on a subject one does not understand proves one to be a fool.
- 4 Seditious pamphlets must be suppressed.
- 5 No one can be expected to remember dates and understand constitutional history.
- 6 Only veteran soldiers should be brought into the field of battle.
- 7 A truly wise man may justly be considered happy.
- 8 Only people who are interested in major sporting events enjoy strenuous physical exercise.
- 9 Only those already engaged in active military service need a boost to their morale.
- 10 Overlooking the need to ensure that your car is in satisfactory mechanical order proves you to be an inexperienced driver.
- 11 Only male teachers with some experience of difficult classes should be put in charge of a difficult class.
- 12 People who take to heart the teachings of the great religious and spiritual teachers of mankind can be entrusted with the role of judge.
- 13 No one can be expected to enjoy both Cottar's Curd and Camembert.

(j) Results of the d method item analysis of VRT Section B.

158 M-Y subjects

Item	R	m(R)	E	m(E)	O	m(O)	U	m(U)	W	m(W)	d	D
5	144	37.7	14 putting Yes 10 putting No(i) 2 putting No(iii) 1 putting No(-) 1	20.4 21.3 27.0 12.0 6.0					14	20.4	17.3	1.24
6	95	41.5	58 putting No(i) 10 putting No(ii) 9 putting No(iii) 9 putting No(iv) 28 putting No(-) 2	28.3 30.2 28.9 21.1 30.0 24.0	5	24.6			63	28.0	13.5	0.97
7	98	39.3	60 putting Yes 16 putting No(i) 19 putting No(iii) 18 putting No(iv) 2 putting No(-) 5	31.0 28.0 34.0 30.1 18.0 38.0					60	31.0	8.3	0.59
8	93	39.8	63 putting Yes 24 putting No(i) 12 putting No(ii) 15 putting No(iii) 9 putting No(-) 3	30.6 35.5 28.0 22.9 33.4 32.3	2	38.5			65	30.8	9.0	0.64
9	117	40.5	40 putting No(i) 23 putting No(ii) 3 putting No(iii) 13	23.1 22.9 26.0 23.0	1	47.0			41	23.7	16.3	1.07

Item	R	m(R)	E	m(E)	O	m(O)	U	m(U)	W	m(W)	d	D
9			putting No(iv) 1	22.0								
10	100	41.9	58 putting Yes 36 putting No(ii) 3 putting No(iii) 3 putting No(iv) 16	26.3 26.8 13.7 25.6 27.4					58	26.3	15.6	1.11
11	92	43.6	65 putting Yes 38 putting No(i) 13 putting No(ii) 3 putting No(iii) 10 putting No(-) 1	25.8 28.4 21.9 23.3 21.3 29.0	1	25.0			66	25.8	17.8	1.28
12	103	40.7	53 putting No(i) 29 putting No(ii) 9 putting No(iii) 1 putting No(iv) 13 putting No(-) 1	27.3 28.4 14.9 25.0 35.1 6.0	2	38.5			55	27.7	13.0	0.93
13	116	40.4	41 putting No(i) 12 putting No(ii) 20 putting No(iii) 3 putting No(iv) 3 putting No(-) 3	23.9 27.1 23.2 17.7 28.0 18.3	1	37.0			42	24.2	16.2	1.16
14	47	45.3	108 putting Yes 81 putting No(i) 8 putting No(ii) 3 putting No(iv) 16	31.9 34.7 23.3 14.7 25.7	2	39.5	1	51.0	111	32.2	13.1	0.94

Item	R	m(R)	E	m(E)	O	m(O)	U	m(U)	W	m(W)	d	D
15A	28	46.1	120 putting Yes 104 putting No(i) 8 putting No(ii) 1 putting No(iv) 2 putting No(-) 5	34.3 33.7 43.8 22.0 14.0 41.0	1	51.0	9	28.1	130	34.0	12.1	0.86
15B	28	46.3	112 putting Yes 35 putting No(i) 10 putting No(iii) 8 putting No(iv) 38 putting No(-) 21	35.2 36.1 29.7 31.8 37.6 33.0			18	27.0	130	34.0	12.3	0.88
<u>160 A-L subjects</u>												
5	158	43.6	2 (both put Yes)	27.5					2	27.5	16.1	1.39
6	121	45.7	35 putting No(i) 9 putting No(ii) 5 putting No(iii) 2 putting No(iv) 19	36.3 34.0 38.8 47.5 35.6	4	32.8			39	36.0	9.7	0.83
7	97	46.1	62 putting Yes 15 putting No(i) 30 putting No(iii) 12 putting No(iv) 2 putting No(-) 3	39.0 35.2 40.8 39.7 38.0 37.3	1	43.0			63	39.1	7.0	0.60
8	123	46.2	37 putting Yes 16	34.1 30.5					37	34.1	12.1	1.04

Item	R	m(R)	E	m(E)	O	m(O)	U	m(U)	W	m(W)	d	D
14			putting No(iv) 3	32.0								
15A	46	50.6	111 putting Yes 92 putting No(i) 9 putting No(ii) 2 putting No(iv) 3 putting No(-) 5	40.8 39.5 49.6 41.5 45.3 44.6	1	33.0	2	25.0	114	40.4	10.2	0.88
15B	57	50.4	97 putting Yes 27 putting No(i) 12 putting No(iii) 11 putting No(iv) 28 putting No(-) 19	40.3 43.2 42.1 38.8 39.2 37.5			6	26.5	103	40.3	10.1	0.87

The categories of response used in the above analysis are as follows:-

- R correct responses
 - E cases where a subject puts an answer but a wrong one
 - O cases where a subject puts no answer to this item but does put an answer to a later item
 - U cases where a subject puts no answer to this item or to any later item
- $W = E + O + U$

$$d = m(R) - m(W)$$

$D = \frac{d}{\sigma}$, where σ is the root of the unbiased estimate of the variance of the population.

(k) Inventive-response answers of 45 subjects to item 7 of the VRT Section B (the item for which none of the reasons given in the published version of the test makes the essential point). These are reproduced here to illustrate the difficulties which would face the marker if we were to follow Anstey's suggestion that the VRT 'might have been even better if cast entirely in inventive-response form'.

The original purpose of the exercise was to obtain empirical evidence about the kind of flaw undergraduate subjects are most prone to see in an argument and, to this end, these members of an Ordinary Psychology class were told that there was 'something wrong' with the six invalid arguments of the VRT Section B and asked to say what that something is. In the following list I have placed first the 17 responses which include the essential point - that the conclusion would follow only if the second premise meant or implied that all ten-pound householders (hence forth 'TPH's') vote for Mr. B. Next come 6 problematic cases of which the first 4 probably do contain the important point. Of the remainder, the first 15 make the rather lame comment that the conclusion does not follow from the premises while the other 7 are, in my view, unacceptable for a variety of other reasons.

'Only Whigs vote for Mr. B. All who vote for Mr. B. are ten-pound householders. Therefore only Whigs are ten-pound householders.'

1 There may be Whigs who don't vote for Mr. B. and also TPH's who don't vote for Mr. B.

2 Some Whigs might not vote at all, or might vote for someone else. The fact that anyone who does vote for him is a Whig doesn't exclude other Whigs who don't vote. The same argument holds for TPH's - as with Whig statements it isn't reversible.

3 All who vote for Mr. B. are TPH's does not mean that everyone who is a TPH votes for Mr. B. Therefore last statement illogical. It is possible for people other than Whigs to be TPH's.

4 Only Whigs who are TPH's vote for Mr. B. but it says nothing about every single TPH voting for Mr. B. - there may be TPH's who don't vote for Mr. B. and all the Whigs may not vote for him. Thus those who are not TPH's don't vote for Mr. B. and since every Whig does not necessarily vote for Mr. B, all Whigs are not nece.... Necessary cond. to vote for Mr. B. - TPH. Also Whig necessary condition. No logical conclusion that only Whigs the TPH's since some Whigs and some TPH's may not vote for Mr. B.

5 Only Whigs vote for Mr. B. yes, but there may be TPH's who do not vote Mr. B. and are not Whigs. It may be just a coincidence that all who vote for Mr. B. are TPH's.

6 Some TPH's may not vote for Mr. B. (i.e., may not vote at all) and it is therefore impossible to say only Whigs are TPH's.

7 TPH's don't necessarily vote for Mr. B. Therefore cannot assume they are all Whigs.

8 The argument does not say that all TPH's vote for Mr. B. The only conclusion could be that all Whigs who vote for Mr. B. are TPH's.

- 9 The fact that all who vote for Mr. B are TPH's does not necessarily imply that there are not any TPH's who vote, say, for Mr. A. or Mr. C. who may not be Whig candidates.
- 10 Would have to read only those who vote etc. to make sense.
- 11 Other people than Whigs could be TPH's - instead of all in second sentence 'only those who' should be used.
- 12 But not all TPH's vote for Mr. B.
- 13 Not only Whigs are TPH's but others can be this as well as they do not have to vote for Mr. B.
- 14 Definition of Whig \neq TPH or even if all Whigs were TPH's, it would not mean that this was exclusively their prerogative to be TPH's, i.e., there may be other TPH's who do not vote for Mr. B.
- 15 Some TPH's may not vote for Mr. B. Therefore there may be some who are not Whigs. Also not all Whigs necessarily vote for Mr. B.
- 16 There may be others who do not vote for Mr. B. who are TPH's.
- 17 The argument does not state that all TPH's vote for Mr. B. There might be some TPH's who are not Whigs that do not vote for Mr. B.
- 18 Being a Whig does not stipulate being a TPH, i.e., there may have been some TPH's who voted otherwise.
- 19 The statement implies that only Whigs are TPH's. Whereas it should say that only Whigs who are TPH's vote for Mr. B. as there are many other categories of TPH's.
- 20 There could be some TPH's who abstained from voting, or who were not in the electoral area. These people need not necessarily be Whigs. This statement is a generalisation based on the evidence of one incident.
- 21 The last sentence does not follow, as there could be Tories who are TPH's, even although they do not vote for Mr. B.
- 22 Conclusion implies 'Whigs only vote for Mr. B.', not 'only Whigs' which implies only Whigs out of all political parties, for example, and not out of all voters.
- 23 There may be some Whigs who don't vote and there may be some TPH's who don't vote. Therefore it is not necessary that those who vote for Mr. B. are either Whigs or TPH's.
- 24 All TPH's are not necessarily Whigs.
- 25 It is possible for people other than Whigs to be TPH's.
- 26 There will be TPH's who are not Whigs. Therefore it is wrong to make sweeping generalisations such as have been made here.
- 27 There may be people who are not Whigs and yet are TPH's.
- 28 There are many TPH's who are not Whigs.
- 29 The Whigs who vote for Mr. B. are TPH's. This does not mean that

only Whigs are TPH's, non-Whigs may be TPH's.

30 The conclusion is stating a general principle taken from only one premise.

31 TPH's may be a larger group than Whigs. Whigs only part of TPH's. Therefore it is not only Whigs who are TPH's.

32 Some Whigs may not vote for Mr. B. Some TPH's may not vote for Mr. B. (All this scored out.) Therefore there are some TPH's who aren't Whigs.

33 Not only Whigs that are TPH's.

34 Only Whigs vote for Mr. B. and these are TPH's, but there will be others who, although they qualify as TPH's, will not be Whigs.

35 Non sequitur.

36 Whigs are not necessarily only people who are TPH's. This would be better if one said 'all Whigs etc.' and not 'only Whigs'.

37 There could be others who are TPH's but not Whigs.

38 Does not make sense to assume that only Whigs are TPH's. There are most likely others in this category.

39 Since all who vote for Mr. B. are TPH's and they are only Whigs - it doesn't follow logically that the whole number of Whigs (as opposed to those who vote for Mr. B.) are TPH's.

40 'Whigs' in conclusion should be 'Whigs who vote for Mr. B.' There are Whigs who do not vote for Mr. B. and some of these are probably not TPH's.

41 The last statement should be changed to 'all the Whigs who vote for Mr. B. are TPH's'.

42 Would make better sense if it said 'all Whigs are TPH's'.

43 Conclusion should be 'All Whigs are TPH's'. There could be others who are not Whigs, but are TPH's.

44 Doesn't necessarily follow that only Whigs vote for Mr. B. and because if someone votes for Mr. B. and is a TPH he still might not be a Whig. Who knows who votes for who anyway?

45 Only those who are also TPH's will vote for Mr. B. This need not be all Whigs.

APPENDIX C

GLOSSARY OF LOGICAL TERMS USED IN THIS THESIS

In the traditional Aristotelian logic all simple statements - i.e., all statements which are not combinations of other statements - are assumed to involve a subject-term and a predicate-term. The subject-term of a proposition is the one which refers to the class of things spoken about in that proposition, the predicate-term refers to the class of things to which the subject class stands in a certain relation of inclusion or exclusion, a relation which is expressed with the help of the remaining component of an Aristotelian proposition, the copula.

A proposition or statement is said to have a certain quality - affirmative or negative - and a certain quantity - universal or particular. The intersection of these two dichotomies produces four types of statement, the universal affirmative (A), the universal negative (E), the particular affirmative (I), and the particular negative (O). In their most characteristic English forms these are, respectively:

- A All X's are Y's.
- E No X's are Y's.
- I Some X's are Y's.
- O Some X's are not Y's.

From the truth of a proposition of one of these kinds an immediate inference can sometimes be made about the truth of another proposition. The simplest case would be where the truth of a particular proposition is inferred from the truth of the corresponding universal. A more complicated case is where a proposition is converted, i.e., where the statement to be inferred from it has the order of the subject and predicate terms ^{reversed} without any compensating change in the copula. Thus, if no X's are Y's, then one may legitimately immediately infer that no Y's are X's. On the other hand, conversions of A and O propositions are illicit: 'All Y's are X's' may not be inferred from 'All X's are Y's' nor 'Some Y's are not X's' from 'Some X's are not Y's'.

Mediate inferences recognised in the Aristotelian logic include the sylogism and the sorites. In a valid syllogism there are three propositions and three terms, two propositions (with all three terms represented) being premises and the remaining one the conclusion. The subject-term is called the minor term of the syllogism and represented by the letter 'S'; the predicate terms of the conclusion (the major term) is represented by 'P'; the remaining term of the syllogism, which must appear in both premises but does not appear in the conclusion, is called the middle term ('M'). The premise in which the major term appears is the major premise and conventionally written first, followed, of course, by the premise with the minor term (the 'minor premise'). An example of a valid syllogism would thus be:

All M is P
All S is M
Therefore all S is P.

The example is a syllogism in Barbara, one of a number of mnemonic words used to refer to a syllogism of a certain figure and mood. The figure of a syllogism is the arrangement of terms when the premises are set out in the conventional order, a syllogism in Barbara being in the first of

the four figures recognised, the middle term in others being in the predicate position in both premises (second figure), the subject position in both (third figure) or in predicate position in the major premise and subject position in the minor (fourth figure). The mood of a syllogism is a matter of the quality and quantity of the component propositions, stated in the conventional order. Thus a syllogism in Barbara is in the AAA mood.

A sorites is a conflation of two or more syllogisms 'in which only the final conclusion is stated and the premises are so arranged that any two successive premises contain a common term'. (Stebbing, 1952, p.70).

There are certain 'rules of the syllogism' which make it possible to determine whether the conclusion follows from the premises. The least obvious of these concern the distribution of terms in premises and conclusion. The subject-terms of universal, and the predicate-terms of negative propositions are distributed (i.e., concern all the members of the classes they refer to), all other terms being undistributed. The relevant rules of the syllogism state (1) that the middle term must be distributed in at least one premise, (2) that the major term must be distributed in its (major) premise if it is distributed in the conclusion, and (3) that the minor term must be distributed in its (minor) premise if it is distributed in the conclusion. A person who affirms the conclusion of a syllogism when one or other of these rules or conditions is not met is guilty of committing, respectively, the fallacy of the undistributed middle term, the fallacy of illicit process of the major term, or the fallacy of illicit process of the minor term.

These are all examples of formal fallacies, fallacies, that is, 'which could be detected solely by a knowledge of formal logic' (Sinclair, 1951, p. 85). The ability to detect material fallacies, on the other hand, depends on such extra-logical factors as a knowledge of the context of which the argument is a part ('begging the question', for example) or of the different meanings which one of the terms used in the argument may have ('equivocation', for example).

APPENDIX D

MATERIALS RELATING TO THE REVIEW OF LITERATURE

(a) Chapman and Chapman's Table showing the percentage of their subjects choosing an A, E, I or O conclusion - or none of these (N) - as following from different pairs of premises in syllogisms of different figures (Chapman and Chapman, 1959, p.

Item No.	Prem-ises	Fig-ure	A	E	I	O	N	Item No.	Prem-ises	Fig-ure	A	E	I	O	N
12	AA	II	83	6	3	1	7	5	II	IV	2	3	68	13	15
17	AA	II	82	5	3	1	9	20	II	III	1	5	63	5	26
39	AA	II	77	5	6	1	10	51	II	III	4	5	64	5	23
4	AE	I	3	81	3	5	8	7	IO	III	1	6	13	48	31
23	AE	III	1	85	0	5	8	34	IO	IV	2	5	11	60	22
41	AE	I	1	82	3	6	7	48	IO	I	2	6	10	55	27
8	AI	II	3	7	75	7	8	22	OI	I	1	4	14	59	21
15	AI	IV	3	3	80	6	8	33	OI	III	1	7	15	52	24
46	AI	IV	10	2	74	6	7	44	OI	IV	1	5	11	55	27
13	IA	I	5	5	78	8	5	29	EE	IV	1	57	4	3	36
19	IA	II	5	11	68	7	9	36	EE	II	3	59	5	5	28
42	IA	I	3	4	83	4	7	40	EE	III	2	47	4	7	40
11	AO	III	2	7	14	61	16	30	EO	I	3	24	10	32	32
24	AO	I	1	2	13	76	8	35	EO	II	1	26	9	32	32
52	AO	IV	4	4	10	74	11	47	EO	III	5	25	6	21	44
25	OA	II	0	7	12	64	16	2	OE	IV	2	28	12	24	34
32	OA	IV	3	4	11	70	12	27	OE	I	3	39	5	19	34
43	OA	I	1	6	7	78	8	50	OE	III	3	41	7	19	30
9	IE	I	1	62	6	13	18	3	OO	III	0	8	10	50	31
26	IE	III	2	59	5	16	19	14	OO	IV	0	5	11	60	24
49	IE	IV	2	48	6	24	20	45	OO	II	1	8	11	45	35

(b) The relationship between the validity of the E ('emotionally significant') and nE ('emotionally neutral') arguments in Lefford's study and the truth or falsity of the corresponding conclusions. The conclusions are listed on the next page with the validity (V) or invalidity (I) of the corresponding arguments immediately to the right. To the right again are the views of three judges (myself and two young American women who were not aware of the issue involved) as to whether Lefford's American subjects in 1941 would have regarded these conclusions as definitely true (T), probably true (T?), definitely false (F), probably false (F?). Two other categories were also used: N ('neither') to refer to conclusions of which it does not appear to make sense to say that they are true or false, and CT ('can't tell') for conclusions about which the judge felt completely unsure. The question at issue is, of course, whether the nE conclusions are more often true when the corresponding argument is valid and false when it is invalid than in the case of E conclusions.

<u>nE conclusions</u>		IW	DS	AN
1 All members of the North End Club will vote on Election Day.	V	N	CT	F
2 All geometrical figures which are plane figures are triangles.	I	F	F	F
3 All automobile accidents should be avoided if possible.	V	T	T	T
4 This person is not a college graduate.	I	N	N	N
5 This plane figure, which has three sides and three included angles, is a triangle.	V	T	T	T
6 The weather will change.	I	N	N	CT
7 The laws of science can never be established as absolute, completely definite, and final.	V	CT	F?	F?
8 All collegiate football players are members of Phi Beta Kappa.	I	F	F	F
9 All good scientists should be masters of the laws of logic and scientific method.	V	T	T	T
10 All good men are wise.	I	CT	F?	T?
11 Logic is a branch of mathematics.	V	CT	T	CT
12 Being a vegetarian is the best indication of good health.	I	F	F	CT
13 Da Vinci's paintings obeyed the laws of perspective.	V	T?	T	T
14 The S.S. America has been constructed of materials which are lighter than water.	I	F	CT	CT
15 Philosophers are fallible.	V	T	T	T
16 All fish must be whales.	I	F	F	F
17 No members of the library committee are members of the finance committee.	V	CT	CT	CT
18 Some forms of plant-life are fish	I	F	F	F
19 Spiders are not insects.	V	T	F	F
20 Auto mobiles will not come into greater demand.	I	F	F	F

E conclusions

1 War is an experience ennobling to men to the most exalted degree.	V	F	F?	F?
2 The Teachers' Union is a subversive Communist organisation.	I	F?	CT	T?
3 The German Jews are justifiably repressed by all measures.	V	F	CT	F?
4 The Soviet government does not want peace.	I	T?	CT	T?
5 The rearmament measures in which we are engaged are wrong.	V	F?	CT	T?
6 This country is not going to war.	I	CT	CT	F
7 Marriages between peoples of different races, as between Negroes and Whites or Jew and Aryan, are bad and undesirable.	V	CT	T?	T?
8 The European war is opposed to our national prosperity and welfare.	I	CT	T	T
9 The poor and unemployed should not be allowed to survive.	V	F	F	F?
10 All trade unionists are communists.	I	F?	CT	T?
11 War should be shunned at all costs.	V	F?	F	F
12 A consequence of communism is, in effect, the enslavement of the working man.	I	T?	T?	T
13 Birth control should be prohibited.	V	CT	T?	T
14 The Russian government is really acting in the best interests of the Russian people.	I	CT	T?	T

Lefford's E conclusions (contd.)

15	Wars are much to be desired.	V	F	F	F
16	All C.I.O. leaders are agents for Communism.	I	F?	CT	T?
17	The existence of God is not real.	V	F?	CT	F
18	Some dictators are without personal ambition.	I	F?	F?	F
19	No peace settlement can be a lasting one.	V	F	F?	F?
20	The annexation of Austria was not rightful.	I	T	T	T

(c) The conclusions of Thistlethwaite's 'anti-Negro' arguments. The validity or otherwise of the corresponding arguments is once again shown at the right hand side. The purpose to be served in this case is illustrative only: Thistlethwaite says that he chose conclusions which would conflict with a correct view of the related arguments in subjects with anti-Negro sentiments.

1	Negroes are best fitted for skilled work or for positions of responsibility.	V
2	The plight of the Negroes is the fault of bigots or of Fascists.	V
3	Negroes who are educated should be admitted to the better hotels and apartments.	V
4	Negroes are not the equals of whites in intelligence.	I
5	The Negro is by nature lazy and superstitious.	I
6	Negroes should feel no hesitation about dating white girls who like them.	V
7	Radical agitators are responsible for the talk about putting Negroes on the same level as Whites.	I
8	If Negroes should be prevented from having too much contact with Whites then it is best to keep Negroes in their own districts and schools.	I
9	If peaceful industrial relations are desirable, then it is a mistake to have Negroes for foreman and leaders over Whites.	I
10	The Negroid group does not represent an advanced and mature race.	I
11	Negroes should be assisted as much as possible in improving their social and economic status.	V
12	No Negro should submit to social conventions without thinking for himself.	V

APPENDIX E

PRELIMINARY DATA RELATING TO THE PR AND GR GROUPS

(a) AAA and VRT, Section B, T-scale scores and Class Exam percentage marks. The AAA scores, as explained in the text, were based on performance on Section A of the VRT, on the CMT (adjusted to allow for the anti-Science bias) and on the Higher grade of the S.C.E. weighted in the proportions 1 : 3 : 4. T-scale scores were calculated to facilitate the selection of subjects for PR and GR groups. In the following table subjects are set out in 'pairs'.

<u>P R g r o u p</u>						<u>G R g r o u p</u>					
Sub- ject	AAA	VRT B	Diff- erence	Class 1st	Exams 2nd	Sub- ject	AAA	VRT B	Diff- erence	Class 1st	Exams 2nd
JA	58	39	-19	63	56	JH	58	61	+3	58	50
EA	63	41	-22	71	62	RM	60	58	-2	48	48
NB	59	46	-13	50	62	HMcL	62	64	+2	42	50
AB	67	49	-18	48	44	AMcD	68	69	+1	56	64
NC	52	40	-12	68	59	JW	50	52	+2	44	56
MC	57	45	-12	71	49	KH	56	57	+1	72	72
LD	65	47	-18	57	54	EW	62	67	+5	52	38
GD	53	36	-17	36	41	JW	56	59	+3	43	40
EF	50	37	-13	63	41	DJ	55	56	+1	56	48
IG	69	45	-24	72	69	BP	68	62	-6	68	62
MH	61	39	-22	46	52	MA	61	61	0	64	62
JH	59	48	-11	54	42	RW	57	58	+1	46	44
JJ	63	45	-18	73	66	MO	60	59	-1	51	52
RJ	59	43	-16	60	57	MC	57	58	+1	56	40
EL	65	45	-20	72	51	GC	64	60	-4	--	--
IMcA	66	49	-17	72	71	WP	66	68	+2	63	65
BMcC	56	44	-12	63	52	EMcL	54	54	0	58	57
HM	61	41	-20	--	--	RH	66	63	+3	67	42
IR	51	29	-22	65	53	RC	51	54	+3	59	55
JS	70	46	-24	78	65	JC	64	65	+1	36	48
PW	55	43	-12	60	55	RD	54	55	+1	52	55
RW	59	39	-20	38	42	MG	58	60	+2	50	36

(b) Responses of the two groups to some items in the questionnaire. The questionnaire is reproduced in full in Appendix B and I have simply indicated the various items by means of their number and a shorthand version of their contents.

<u>Item</u> <u>No.</u>		<u>PR</u>	<u>GR</u>	<u>Item</u> <u>No.</u>	<u>PR</u>	<u>GR</u>
2. Sex:	Male	8	11	3. Age:	18	4 7
	Female	13	10		19	7 6
5. Faculty:	Arts	17	17	20	9 5	
	Science	4	4	21	- 2	
				21+	1 1	

<u>Item No.</u>		<u>PR</u>	<u>GR</u>	<u>Item No.</u>		<u>PR</u>	<u>GR</u>	
6.	Year of study:			31.	Spent on? Section A	3	2	
		1st	2		3	Section B	5	3
		2nd	11		11	Both	9	1
		3rd	8		6			
	4th	-	1					
10.	VRT good test?			32.	Time limit upset Yes	13	8	
		Yes	10		4	No	8	13
		No	9	12				
	D.K.	2	5	33.	Task clear? Very	6	8	
					Quite	11	12	
					No	4	1	
14.	Honours?			34.	Guess? Often	3	-	
		Yes	12		15	Occasionally	11	13
		No	8		6	Never	7	8
	D.K.	1	-					
18.	Good at spotting fallacies?			35.	Reasons first Always	2	-	
		Very	1		-	Often	2	2
		Quite	8		15	Occasionally	13	11
		Not very	11		4	Never	4	8
		Hopeless	1		1			
	D.K.	-	1					
19.	Others better?			36.	CMT more/less enjoyable? More	14	12	
		Yes	18		17	Less	2	6
		No	1		-	Same	5	3
		D.K.	2		3			
	Other	-	1					
20.	Logic classes?			43.	One Form easier? Yes	13	11	
		School	-		-	No	8	10
		University	7		6			
	Neither	14	15					
21.	Read logic books?				Second Form? Yes	13	11	
		Yes	2	3	No	3	2	
		No	18	18				
	?	1	-					
27.	Logic of use?				Form W, second	3	7	
		Yes	4	4	Form V, second	7	2	
		No	2	1	Form W, first	3	-	
		D.K.	1	-	Form V, first	-	2	
		Doesn't apply	13	16				
	No answer	1	-					
28.	Logic aids?			44.	Task clearer/practice	7	8	
		Symbols	4		1	Less nervous	2	3
		Venn	2		1	Others (inc. novelty)	4(1)	1(1)
		Conversion/obversion	1		1	None given	1	1
		Only	3		2			
		None	3		1			
	Doesn't apply	12	16					
29.	More time?			47.	Special disability mentioned	9	9	
		Yes	17		6			
		No	3		14			
	D.K.	1	1					
30.	How much?			48.	One test better? CMT	13	6	
		A few mins.	-		2	VRT	4	7
		10 mins.	2		-	No	4	7
		10-15 mins.	1		-	D.K.	-	1
		15 mins.	8		4			
		30 mins.	3		-			
		60 mins.	1		-			
	Not sure	2	-					
30.	How much?			49.	One test worse? CMT	3	7	
		A few mins.	-		2	VRT	12	4
		10 mins.	2		-	No	6	8
		10-15 mins.	1		-	D.K.	-	2

APPENDIX F

MATERIALS AND RAW DATA RELATING TO THE 5TS EXPERIMENT

(a) Statements and diagrammatic card arrays. Those used in the fifth series are reproduced in the text. Accordingly, only the first four arrays for each statement-type are given below.

Ah (1) R t M p 3 2 7 8

If a card has a capital letter on one side it always has an odd number on the other.

Ah (2) + Δ \square Δ e R t m

If a card has a cross on one side it always has a small letter on the other.

Ah (3) α β δ e 3 5 7 9

If a card has a Greek letter on one side it always has an even number on the other.

Ah (4) + \square Δ \square δ p t α

If a card has a square on one side it always has a Greek letter on the other.

Ad (1) \square \diamond + + a e K t

Every card has a cross on one side or else a capital letter on the other.

Ad (2) A W T E 8 2 3 4

Every card has a vowel on one side or else an even number on the other.

Ad (3) α a c b Δ \square \diamond +

Every card has a Greek letter on one side or else a square on the other.

Ad (4) P A G X Δ \square Δ Δ

Every card has a consonant on one side or else a triangle on the other.

A (1) K t R d 2 3 4 5

All the cards with a capital letter on one side have an even number of the other.

A (2) Δ Δ Δ \square m t R T

All the cards with a triangle on one side have a capital letter on the other.

A (3) α β a m 7 8 4 2

All the cards with a Greek letter on one side have an even number on the other.

A (4) β π b α \diamond + \square \diamond

All the cards with a Greek letter on one side have a square on the other.

E (1) a α β b 7 3 2 4

No cards with a Greek letter on one side have an odd number on the other.

E (2) α t m e \square + Δ \diamond

No cards with a Roman letter on one side have a square on the other.

E (3) 2 3 4 5 Δ + \square \square

No cards with an even number on one side have a triangle on the other.

E (4) \square \square \square + k s e t

No cards with a square on one side have a consonant on the other.

F (1) r M m T 2 4 7 6

Only cards with a small letter on one side have an even number on the other.

F (2) \square + Δ + L Z p a

Only cards with a cross on one side have a capital letter on the other.

F (3) 3 6 8 2 + \square Δ Δ

Only cards with an even number on one side have a square on the other.

F (4) Δ Δ \square + a e i z

Only cards with a triangle on one side have a vowel on the other.

(b) Response times on successive trials of the 5TS task, totals and ranks. The times are given in tenths of a second to avoid the need for the decimal point. Subjects will be given in their pairs in the order in which they are listed in Appendix E, and this will hold good for all the results presented in the Appendices from this point on. Thus the subject numbers will always refer to the same pair of subjects.

Ah statements

Sub- ject	<u>GR group</u>						<u>PR group</u>						
	1	<u>T r i a l s</u>				5	Total	1	<u>T r i a l s</u>				5
1	471	208	327	249	316	1571	597	185	277	439	224	1722	
2	312	142	123	73	319	969	119	162	256	159	143	839	
3	229	90	128	62	73	582	421	199	320	464	811	2215	
4	109	84	193	148	119	653	97	69	73	86	78	403	
5	124	104	79	85	111	503	177	90	85	99	73	524	
6	168	155	401	151	107	982	251	216	331	172	234	1204	
7	236	155	173	113	165	842	1508	189	197	182	265	2341	
8	328	468	398	542	150	1886	225	100	171	196	122	814	
9	153	104	81	158	102	598	114	125	462	84	98	883	
10	293	516	288	243	398	1738	467	197	116	178	270	1228	
11	86	142	144	130	232	734	133	140	157	107	97	634	
12	117	285	152	193	92	839	75	65	105	74	94	413	
13	136	422	326	128	90	1102	163	194	245	150	90	842	
14	116	133	87	170	100	606	432	222	226	194	201	1275	
15	45	31	35	25	37	173	79	68	86	67	140	440	
16	154	119	144	116	148	681	207	217	157	183	114	878	
17	615	292	487	213	323	1930	239	209	215	135	159	957	
18	115	60	64	65	50	354	92	120	111	102	99	524	
19	683	362	748	575	462	2770	323	389	281	173	153	1319	
20	157	115	76	66	90	504	419	268	159	190	189	1225	
21	376	699	399	221	95	1790	269	492	536	164	169	1630	
22	329	140	147	94	125	835	97	135	57	64	129	482	

Ad statements

1	480	585	281	270	404	2020	494	187	150	163	192	1186
2	510	191	134	144	277	1256	177	162	195	239	191	964
3	150	167	177	88	80	662	422	576	1013	265	725	3001
4	152	135	113	130	108	638	254	152	169	68	132	775
5	84	186	110	81	143	604	439	178	147	144	135	1043
6	442	405	527	129	82	1585	473	235	335	222	599	1864
7	180	151	109	145	139	724	828	327	134	482	172	1943
8	348	806	365	186	386	2091	217	121	125	253	129	845
9	546	141	164	442	130	1423	170	186	140	130	177	803
10	395	246	258	198	176	1273	954	184	259	190	308	1895
11	158	270	142	143	170	883	146	148	378	245	205	1122
12	132	229	223	127	168	879	189	242	174	163	183	951
13	130	229	251	122	126	858	226	225	186	182	154	973
14	185	100	93	130	144	652	377	251	215	303	315	1461
15	65	45	39	49	37	235	98	79	210	113	136	636
16	281	561	190	233	127	1392	418	384	346	171	110	1429
17	272	391	214	324	214	1415	370	231	422	308	294	1625
18	162	162	110	112	77	623	452	210	70	82	95	909
19	486	449	893	269	349	2446	395	385	167	428	210	1585
20	75	216	181	190	194	856	464	161	144	412	326	1507
21	402	465	379	339	368	1953	373	357	463	203	341	1737
22	579	184	145	98	189	1195	152	158	110	108	230	758

Sub- ject	<u>GR group</u>						<u>PR group</u>					
	1	2	3	4	5	Total	1	2	3	4	5	Total
<u>A statements</u>												
1	389	394	277	292	319	1671	352	403	177	671	148	1751
2	134	167	203	262	249	812	188	172	142	167	112	781
3	144	110	124	89	50	517	430	469	441	778	449	2567
4	107	78	84	155	94	518	73	103	119	59	74	428
5	266	78	185	81	87	697	142	76	111	127	107	563
6	438	98	156	91	94	877	305	224	207	281	348	1363
7	318	179	186	197	186	1066	846	443	404	517	187	2397
8	217	876	367	315	241	2016	131	159	95	122	105	612
9	107	77	105	160	120	569	223	143	102	120	218	806
10	355	557	244	299	235	1690	211	87	172	187	113	770
11	196	67	210	118	198	789	170	118	167	87	158	700
12	97	75	154	158	117	601	59	67	145	71	126	468
13	120	87	132	74	85	498	184	229	165	157	128	863
14	79	207	119	62	98	565	181	135	139	150	185	790
15	41	33	33	39	28	174	113	59	108	65	125	470
16	193	269	152	139	146	899	143	225	141	176	194	879
17	534	388	387	348	202	1859	141	151	109	159	175	735
18	55	39	68	49	58	269	88	131	88	82	92	481
19	449	828	472	512	496	2757	705	377	355	381	120	1938
20	65	108	86	78	101	438	291	527	154	86	154	1212
21	316	227	211	198	160	1112	338	225	508	191	134	1396
22	124	115	83	104	98	524	79	96	84	86	89	434

<u>E statements</u>												
1	210	352	354	470	103	1489	321	230	173	205	183	1112
2	277	324	166	241	135	1143	154	138	129	129	100	650
3	194	179	94	215	114	796	264	905	517	449	325	2464
4	142	152	89	87	107	577	95	113	57	152	111	528
5	84	107	61	71	140	463	217	109	70	85	145	626
6	168	139	97	102	184	690	158	144	267	217	191	977
7	232	199	192	117	197	937	574	254	715	192	255	1190
8	442	480	348	233	315	1818	134	136	127	163	127	687
9	105	155	63	85	82	490	217	232	182	112	196	939
10	193	313	180	292	183	1161	417	182	117	153	126	995
11	368	217	191	127	266	1169	271	150	135	249	273	1078
12	100	279	97	116	166	758	113	281	105	131	244	874
13	105	188	130	147	75	645	319	201	198	154	138	1020
14	97	109	159	173	95	633	195	205	196	241	367	1204
15	47	91	48	41	43	270	212	102	70	110	107	601
16	292	214	130	148	102	886	297	188	336	122	190	1133
17	669	368	313	269	614	2233	227	255	115	312	127	1036
18	107	68	72	75	75	397	149	182	209	141	134	815
19	505	557	285	443	249	2039	310	348	440	265	118	1481
20	296	119	100	81	205	801	150	126	110	106	118	610
21	406	221	129	192	158	1106	211	337	221	171	160	1100
22	170	176	134	201	132	813	200	132	45	69	108	554

Sub- ject	<u>GR group</u>						<u>PR group</u>					
	1	T r i a l s			5	Total	1	T r i a l s			5	Total
		2	3	4				2	3	4		
		<u>F statements</u>										
1	829	330	378	350	258	2145	517	502	252	342	301	1914
2	693	331	170	287	109	1590	325	257	141	225	195	1143
3	334	313	175	98	370	1290	287	689	879	1667	225	3747
4	181	143	90	127	94	635	247	87	72	103	155	664
5	168	159	57	109	129	622	154	118	63	83	106	524
6	225	164	119	104	122	734	239	187	214	213	383	1236
7	312	293	194	223	165	1187	780	462	304	248	283	2077
8	522	695	296	389	483	2383	275	175	185	139	215	989
9	100	81	158	110	128	577	203	114	114	100	120	651
10	665	362	359	335	180	1901	689	215	83	101	106	1194
11	297	405	204	323	532	1761	144	130	158	97	110	639
12	181	164	233	186	204	968	98	95	96	187	744	1220
13	215	405	247	207	147	1221	246	386	253	293	220	1398
14	123	71	115	204	130	643	160	201	282	135	414	1192
15	32	35	38	25	56	186	90	125	122	72	132	541
16	194	154	194	198	150	890	247	638	337	185	222	1629
17	902	344	591	355	770	2962	467	97	265	169	390	1388
18	111	87	73	82	128	481	164	160	119	168	129	740
19	715	572	912	259	503	2961	202	219	339	399	146	1305
20	184	117	194	154	122	771	330	295	217	284	235	1361
21	337	290	303	210	239	1379	515	233	343	368	271	1730
22	490	142	107	148	125	1012	105	130	69	75	89	468

(c) Number of trials in which subjects made the appropriate responses to all eight cards. Since there were five trials for each statement-type, subjects scoring 5 made a completely correct response to that statement type. (Cp. Tables 3.3 and 3.4.)

	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	4	1	1	4	1	4	5	4	4	3
2	0	0	0	1	0	1	3	0	3	0
3	0	4	0	4	1	1	2	0	1	0
4	3	5	3	4	5	0	4	0	4	0
5	0	0	0	3	0	0	5	0	5	0
6	0	0	0	0	0	0	1	0	4	0
7	0	5	1	4	0	5	5	5	4	4
8	0	4	1	4	2	0	5	0	4	0
9	0	3	0	4	0	0	0	0	0	0
10	1	4	2	5	3	0	1	0	0	1
11	0	5	0	3	0	0	0	0	3	0
12	0	4	0	3	2	0	2	0	1	1
13	1	5	0	4	0	0	0	0	4	0
14	1	4	0	4	0	0	0	0	0	0
15	0	0	0	0	0	1	3	1	4	1
16	4	5	4	3	4	1	0	0	2	0
17	0	5	0	4	0	0	4	0	5	0
18	0	2	0	5	0	0	0	0	5	0
19	0	5	3	5	3	0	0	0	0	0
20	0	0	0	0	1	0	3	0	0	0
21	0	0	0	3	0	0	1	0	0	0
22	0	5	0	5	0	0	0	0	0	0

(d) 'Method 3' scores over five trials for the five statement-types, intergroup differences and difference ranks.

Sub- ject	<u>GR_group</u>					<u>PR_group</u>				
	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	19	12	16	19	13	19	20	18	19	16
2	11	8	11	12	9	13	18	10	18	10
3	11	19	13	18	13	16	13	12	14	13
4	18	20	18	19	20	10	18	11	19	8
5	8	3	10	17	7	10	20	10	20	10
6	2	3	1	2	3	9	10	9	19	9
7	15	20	16	19	13	20	20	20	18	18
8	11	16	15	17	13	14	20	15	19	10
9	14	16	12	19	11	10	0	10	8	9
10	15	18	15	20	18	3	5	9	13	10
11	13	20	15	17	11	10	6	10	18	9
12	10	16	12	17	17	15	11	15	16	16
13	16	20	14	19	15	10	10	11	17	12
14	15	18	13	18	10	7	5	6	7	4
15	15	10	15	14	5	12	12	12	18	12
16	19	20	19	18	16	15	10	14	16	10
17	15	20	15	19	12	10	19	10	20	9
18	15	15	15	20	13	15	10	15	20	10
19	13	20	18	20	16	14	8	7	14	5
20	15	5	15	14	14	15	16	15	15	9
21	8	4	9	17	8	6	13	5	12	3
22	15	20	15	20	15	12	0	13	13	6

	<u>Intergroup Differences</u>					<u>Difference Ranks</u>				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	0	-8	-2	0	-3		9	4	$1\frac{1}{2}$	$7\frac{1}{2}$
2	-2	-10	+1	-6	-1	5	12	$1\frac{1}{2}$	$15\frac{1}{2}$	$1\frac{1}{2}$
3	-5	+6	+1	+4	0	$14\frac{1}{2}$	7	$1\frac{1}{2}$	$12\frac{1}{2}$	
4	+8	+2	+7	0	+12	$14\frac{1}{2}$	$2\frac{1}{2}$	$15\frac{1}{2}$	$1\frac{1}{2}$	21
5	-2	-17	0	-2	-3	5	20	$2\frac{1}{2}$	$9\frac{1}{2}$	$7\frac{1}{2}$
6	-7	-7	-8	-17	-6	18	8	17	21	15
7	-5	0	-4	+1	-5	$14\frac{1}{2}$		$9\frac{1}{2}$	5	12
8	-3	-4	0	-2	+3	$8\frac{1}{2}$	4	$2\frac{1}{2}$	$9\frac{1}{2}$	$7\frac{1}{2}$
9	+4	+16	+2	+11	+2	$11\frac{1}{2}$	19	4	$19\frac{1}{2}$	$3\frac{1}{2}$
10	+12	+13	+6	+7	+8	21	$16\frac{1}{2}$	14	$17\frac{1}{2}$	18
11	+3	+14	+5	-1	+2	$8\frac{1}{2}$	18	12	5	$3\frac{1}{2}$
12	-5	+5	-3	+1	+1	$14\frac{1}{2}$	$5\frac{1}{2}$	7	5	$1\frac{1}{2}$
13	+6	+10	+3	+2	+3	17	12	7	$9\frac{1}{2}$	$7\frac{1}{2}$
14	+8	+13	+7	+11	+6	$19\frac{1}{2}$	$16\frac{1}{2}$	$15\frac{1}{2}$	$19\frac{1}{2}$	15
15	+3	-2	+3	-4	-7	$8\frac{1}{2}$	$2\frac{1}{2}$	7	$12\frac{1}{2}$	17
16	+4	+10	+5	+2	+6	$11\frac{1}{2}$	12	12	$9\frac{1}{2}$	15
17	+5	+1	+5	-1	+3	$14\frac{1}{2}$	1	12	5	$7\frac{1}{2}$
18	0	+5	0	0	+3	$1\frac{1}{2}$	$5\frac{1}{2}$	$2\frac{1}{2}$		$7\frac{1}{2}$
19	-1	+12	+11	+6	+11	3	15	18	$15\frac{1}{2}$	20
20	0	-11	0	-1	+5	$1\frac{1}{2}$	14	$2\frac{1}{2}$	5	12
21	+2	-9	+4	+5	+5	5	10	$9\frac{1}{2}$	14	12
22	+3	+20	+2	+7	+9	$8\frac{1}{2}$	21	4	$12\frac{1}{2}$	19

(e) 'Method 4' scores over five trials for the five statement-types, intergroup differences and difference ranks.

Sub- ject	<u>GR group</u>					<u>PR group</u>				
	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	3	16	10	3	16	2	0	4	2	9
2	15	24	18	14	18	14	2	18	3	19
3	16	1	15	4	15	9	13	14	11	15
4	2	0	3	1	0	12	3	18	3	22
5	20	27	19	5	18	20	0	20	0	20
6	20	22	23	21	19	17	17	18	1	19
7	7	0	6	2	15	0	0	0	4	4
8	16	8	10	6	11	14	0	11	2	23
9	13	9	17	1	16	17	40	20	24	21
10	8	2	9	0	3	23	24	15	13	14
11	16	0	11	3	19	20	26	20	3	22
12	19	8	16	5	9	12	19	11	7	7
13	10	0	13	2	12	20	18	18	6	18
14	12	4	14	4	20	19	18	18	17	28
15	11	18	11	12	31	13	16	16	3	15
16	4	0	1	2	8	5	18	9	6	22
17	7	0	9	3	16	20	1	20	0	23
18	12	8	11	0	17	12	22	11	0	20
19	10	0	5	0	8	14	21	19	10	29
20	12	26	11	11	12	12	8	11	9	22
21	17	27	19	3	17	20	12	23	11	19
22	12	0	11	0	8	16	32	13	12	22

	<u>Intergroup differences</u>					<u>Difference Ranks</u>				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	-1	+16	+6	+1	+7	4	11	15	3	11½
2	-1	+22	0	+11	-1	4	16½		17	3
3	-7	-12	+1	-7	0	15½	7	4	13	1½
4	+18	-3	-15	-2	-22	21	3	21	5½	22
5	0	+27	-1	+5	-2		19	3	12	5
6	-3	+5	+5	+20	0	8½	4	11½	21	1½
7	-7	0	+6	-2	+11	15½		15	5½	15½
8	-2	+8	-1	+4	-12	6½	5	4	10	17
9	+4	-31	-3	-23	-5	11½	20	7	22	9
10	+15	-22	-6	-13	-11	20	16½	15	19½	15½
11	+4	-26	-9	0	-3	11½	18	18	1½	7½
12	-7	-11	+5	-2	+2	15½	6	11½	5½	5
13	+10	-18	-5	-4	-6	18	13	11½	10	10
14	+7	-14	-4	-13	-8	15½	8½	8½	19½	13
15	+2	+2	-5	+9	+16	6½	2	11½	15	20
16	+1	-18	-8	-4	-14	4	13	17	10	18½
17	+13	-1	-11	+3	-7	19	1	19	8	11½
18	0	-14	0	0	-3	1½	8½	1½	1½	7½
19	+4	-21	-14	-10	-21	11½	15	20	16	21
20	0	+18	0	+2	-10	1½	13	1½	5½	14
21	+3	+15	-4	-8	-2	8½	10	8½	14	5
22	+4	-32	-2	-12	-14	11½	2	6	18	18½

(f) 'Method 3' scores over five trials on the two half-sets separately. I have not tested the significance of the intergroup differences where they appeared, on inspection, to be very small. Difference ranks for such cases are not given.

Left hand half-set

Sub- ject	<u>GR group</u>					<u>PR group</u>				
	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	10	6	10	10	6	10	10	10	10	7
2	8	4	7	6	6	10	8	9	10	6
3	8	10	10	10	6	10	6	9	10	4
4	10	10	10	9	10	10	9	10	10	1
5	7	1	9	8	0	10	10	10	10	0
6	1	1	0	0	2	8	5	8	10	4
7	10	10	10	10	5	10	10	10	8	10
8	10	8	10	9	7	9	10	10	10	5
9	10	8	10	9	5	9	0	10	8	0
10	10	9	10	10	8	2	3	4	6	5
11	10	10	10	10	9	10	3	10	9	0
12	8	8	10	10	9	10	6	10	10	9
13	9	10	10	10	5	10	5	10	8	4
14	10	8	10	10	0	4	2	3	4	1
15	10	5	10	9	0	10	6	9	8	3
16	10	10	9	8	8	10	5	9	8	3
17	10	10	9	10	5	9	9	8	10	3
18	10	7	10	10	5	10	5	10	10	0
19	9	10	10	10	8	10	3	6	7	1
20	10	3	10	9	8	10	8	10	10	5
21	8	2	9	9	2	6	6	5	5	2
22	10	10	10	10	5	8	0	8	7	1

Intergroup Differences

Difference Ranks

1	0	-4	0	0	-1	9	6
2	-2	-4	-2	-4	0	9	2½
3	-2	+4	+1	0	+2	9	9½
4	0	+1	0	-1	+9	2	21½
5	-3	-9	-1	-2	0	20	2½
6	-7	-4	-8	-10	-2	9	9½
7	0	0	0	+2	-5		17½
8	+1	-2	0	-1	+2	5	9½
9	+1	+8	0	+1	+5	19	17½
10	+8	+6	+6	+4	+3	15½	12½
11	0	+7	0	+1	+9	17½	21½
12	-2	+2	0	0	0	5	2½
13	-1	+5	0	+2	+1	13	6
14	+6	+6	+7	+6	-1	15½	6
15	0	-1	+1	+1	-3	2	12½
16	0	+5	0	0	+5	13	17½
17	+1	+1	+1	0	+2	2	9½
18	0	+2	0	0	+5	5	17½
19	-1	+7	+4	+3	+7	17½	20
20	0	-5	0	+1	+3	13	14½
21	+2	+4	+4	+4	0	9	2½
22	+2	+10	+2	+3	+4	21	14½

Right hand half-set

Sub- ject	<u>GR_group</u>					<u>PR_group</u>				
	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	9	6	6	9	7	9	10	8	9	9
2	3	4	4	6	3	3	10	1	8	4
3	3	9	3	8	7	6	7	3	4	9
4	8	10	8	10	10	0	9	1	9	7
5	1	2	1	9	7	0	10	0	10	10
6	1	2	1	2	1	1	5	1	9	5
7	5	10	6	9	8	10	10	10	10	8
8	1	8	5	8	6	5	10	5	9	5
9	4	8	2	10	6	1	0	0	0	9
10	5	9	5	10	10	1	2	5	7	5
11	3	10	5	7	2	0	3	0	9	9
12	2	8	2	7	8	5	5	5	6	7
13	7	10	4	9	10	0	5	1	9	8
14	5	10	3	8	10	3	3	3	3	3
15	5	5	5	5	5	2	6	3	10	9
16	9	10	10	10	8	5	5	5	8	7
17	5	10	6	9	7	1	10	2	10	6
18	5	8	5	10	7	5	5	5	10	10
19	4	10	8	10	8	4	5	1	7	4
20	5	2	5	5	6	5	8	5	5	4
21	0	2	0	8	6	0	7	0	7	1
22	5	10	5	10	10	4	0	5	6	5

Intergroup Differences

1	0	-4	-2	0	-2
2	0	-6	+3	-2	-1
3	-3	+2	0	+4	-2
4	+8	+1	+7	+1	+3
5	+1	-8	+1	-1	-3
6	0	-3	0	-7	-4
7	-5	0	-4	-1	0
8	-4	-2	0	-1	+1
9	+3	+8	+2	+10	-3
10	+4	+7	0	+3	+5
11	+3	+7	+5	-2	-7
12	+3	+3	-3	+1	+1
13	+7	+5	+3	0	+2
14	+2	+7	0	-5	+7
15	+3	-1	+2	-5	-4
16	+4	+5	+5	+2	+1
17	+4	0	+4	-1	+1
18	0	+3	0	0	-2
19	0	+5	+7	+3	+4
20	0	-6	0	0	+1
21	0	-1	0	+1	+5
22	+1	+10	0	+4	+5

Difference Ranks

	11	11
3 $\frac{1}{2}$	15 $\frac{1}{2}$	14
12	6 $\frac{1}{2}$	
21	4	20 $\frac{1}{2}$
7 $\frac{1}{2}$	20 $\frac{1}{2}$	9 $\frac{1}{2}$
3 $\frac{1}{2}$	9	4 $\frac{1}{2}$
19	1 $\frac{1}{2}$	16 $\frac{1}{2}$
16 $\frac{1}{2}$	6 $\frac{1}{2}$	4 $\frac{1}{2}$
12	20 $\frac{1}{2}$	11
16 $\frac{1}{2}$	18	4 $\frac{1}{2}$
12	18	18 $\frac{1}{2}$
12	9	14
20	13	14
9	18	4 $\frac{1}{2}$
12	4	11
16 $\frac{1}{2}$	13	18 $\frac{1}{2}$
16 $\frac{1}{2}$	1 $\frac{1}{2}$	16 $\frac{1}{2}$
3 $\frac{1}{2}$	9	4 $\frac{1}{2}$
3 $\frac{1}{2}$	13	20 $\frac{1}{2}$
3 $\frac{1}{2}$	15 $\frac{1}{2}$	4 $\frac{1}{2}$
3 $\frac{1}{2}$	4	4 $\frac{1}{2}$
7 $\frac{1}{2}$	22	4 $\frac{1}{2}$

(g) 'Method 4' scores over five trials on the two half-sets separately. I have not tested the significance of intergroup differences where they appeared, on inspection, to be very small. Difference ranks for such cases are not given.

Left hand half-set

Sub- ject	<u>GR_group</u>					<u>PR_group</u>				
	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	0	8	0	0	10	0	0	0	0	7
2	2	12	6	7	6	0	2	1	0	8
3	2	0	0	0	5	0	7	1	1	13
4	0	0	0	1	0	0	2	0	0	15
5	3	13	1	3	15	0	0	0	0	20
6	9	11	12	12	8	2	8	3	0	13
7	0	0	0	0	12	0	0	0	4	0
8	0	4	0	2	3	2	0	0	0	12
9	0	5	0	1	9	1	20	0	4	20
10	0	1	0	0	3	9	10	8	7	6
11	0	0	0	0	2	0	13	0	1	20
12	4	4	0	0	1	0	8	0	0	1
13	3	0	0	0	12	0	9	0	4	14
14	0	4	0	0	20	9	9	8	7	16
15	0	9	0	1	20	0	8	2	3	13
16	0	0	1	2	4	0	9	2	3	15
17	0	0	1	0	12	2	1	4	0	16
18	0	6	0	0	12	0	11	0	0	20
19	1	0	0	0	4	0	11	6	4	17
20	0	12	0	2	4	0	4	0	0	11
21	2	12	1	1	12	5	6	7	7	8
22	0	0	0	0	8	3	16	2	5	11

Intergroup Differences

Difference Ranks

1	0	+8	0	0	+3	11½	4
2	+2	+10	+5	+7	-2	16	1½
3	+2	-7	-1	-1	-8	10	13½
4	0	-2	0	+1	-15	3	20
5	+3	+13	+1	+3	-5	18½	9½
6	+7	+3	+9	+12	-5	4	9½
7	0	0	0	-4	+12		18
8	-2	+4	0	+2	-9	5½	15
9	-1	-15	0	-3	-11	20	16½
10	-9	-9	-8	-7	-3	14	4
11	0	-13	0	-1	-18	18½	21
12	+4	-4	0	0	0	5½	
13	+3	-9	0	-4	-2	14	1½
14	-9	-5	-8	-7	+4	7½	7
15	0	+1	-2	+2	+7	1½	11½
16	0	-9	-1	-1	-11	14	16½
17	-2	-1	-3	0	-4	1½	7
18	0	-5	0	0	-8	7½	13½
19	+1	-11	-6	-4	-13	17	19
20	0	+8	0	-2	-7	11½	11½
21	-3	+6	-6	-6	+4	9	7
22	-3	-16	-2	-5	-3	21	4

Right hand half-set

Sub- ject	<u>GR_group</u>					<u>PR_group</u>				
	Statement Type					Statement Type				
	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	3	8	10	3	6	2	0	4	2	2
2	13	12	12	7	12	14	0	17	3	11
3	14	1	15	4	7	9	6	13	10	2
4	2	0	3	0	0	20	1	18	3	7
5	17	14	18	2	3	20	0	20	0	0
6	11	11	11	9	11	15	9	15	1	6
7	7	0	6	2	3	0	0	0	0	4
8	16	4	10	4	8	12	0	11	2	11
9	13	4	17	0	7	16	20	20	20	1
10	8	1	9	0	0	14	14	7	6	8
11	16	0	11	3	17	20	13	20	2	2
12	15	4	16	5	8	12	11	11	7	6
13	7	0	13	2	0	20	9	18	2	4
14	12	0	14	4	0	10	9	10	10	12
15	11	9	11	11	11	13	8	14	0	2
16	4	0	0	0	4	5	9	7	3	7
17	7	0	8	3	4	18	0	16	0	7
18	12	2	11	0	5	12	11	11	0	0
19	9	0	5	0	4	14	10	13	6	12
20	12	14	11	9	8	12	4	11	9	11
21	15	15	18	2	5	15	6	16	4	11
22	12	0	11	0	0	13	16	11	7	11

Intergroup DifferencesDifference Ranks

	Ah	Ad	A	E	F	Ah	Ad	A	E	F
1	+1	+8	+6	+1	+4	4 $\frac{1}{2}$	9	15 $\frac{1}{2}$		
2	-1	+12	-5	+4	+1	4 $\frac{1}{2}$	17	13		
3	+5	-5	+2	-6	+5	15 $\frac{1}{2}$	7	5 $\frac{1}{2}$		
4	-18	-1	-15	-3	-7	21	3 $\frac{1}{2}$	21		
5	-3	+14	-2	+2	+3	10	20	5 $\frac{1}{2}$		
6	-4	+2	-4	+8	+5	13	5	10 $\frac{1}{2}$		
7	+7	0	+6	+1	+4	18	1 $\frac{1}{2}$	15 $\frac{1}{2}$		
8	+4	+4	-1	+2	-3	13	6	3		
9	-3	-16	-3	-20	+6	10	21 $\frac{1}{2}$	8 $\frac{1}{2}$		
10	-6	-13	+2	-6	-8	17	18 $\frac{1}{2}$	5 $\frac{1}{2}$		
11	-4	-13	-9	+1	+15	13	18 $\frac{1}{2}$	20		
12	+3	-7	+5	-2	+2	10	8	13		
13	-13	-9	-5	0	-4	20	12	13		
14	+2	-9	+4	-6	-12	7 $\frac{1}{2}$	12	10 $\frac{1}{2}$		
15	-2	+1	-3	+11	+9	7 $\frac{1}{2}$	3 $\frac{1}{2}$	8 $\frac{1}{2}$		
16	-1	-9	-7	-3	-3	4 $\frac{1}{2}$	12	17		
17	-11	0	-8	+3	-3	19	1 $\frac{1}{2}$	18 $\frac{1}{2}$		
18	0	-9	0	0	+5		12			
19	-5	-10	-8	-6	-8	15 $\frac{1}{2}$	15 $\frac{1}{2}$	18 $\frac{1}{2}$		
20	0	+10	0	0	-3	1 $\frac{1}{2}$	15 $\frac{1}{2}$	1 $\frac{1}{2}$		
21	0	+9	+2	-2	-6	1 $\frac{1}{2}$	12	5 $\frac{1}{2}$		
22	-1	-16	0	-7	-11	4 $\frac{1}{2}$	21 $\frac{1}{2}$	1 $\frac{1}{2}$		

(h) Errors of commission and errors of omission in the right hand half-sets of Ah and A and the left hand half-set of F.

Errors of Commission

Sub- ject	<u>GR group</u>			<u>PR group</u>			<u>Differences</u>			<u>Ranks</u>		
	Ah	A	F	Ah	A	F	Ah	A	F	Ah	A	F
1	3	7	10	0	2	5	+3	+5	+5	6	14 $\frac{1}{2}$	10 $\frac{1}{2}$
2	2	3	3	6	6	6	-4	-3	-3	10	8	6 $\frac{1}{2}$
3	4	4	4	0	2	8	+4	+2	-4	10	6	8 $\frac{1}{2}$
4	0	0	0	8	7	10	-8	-7	-10	19 $\frac{1}{2}$	19 $\frac{1}{2}$	18 $\frac{1}{2}$
5	5	8	9	8	8	12	-3	0	-3	6	2 $\frac{1}{2}$	6 $\frac{1}{2}$
6	4	5	5	8	9	12	-4	-4	-7	10	11	14
7	7	6	12	0	0	0	+7	+6	+12	16 $\frac{1}{2}$	17 $\frac{1}{2}$	21
8	8	4	1	2	0	12	+6	+4	-11	13 $\frac{1}{2}$	11	20
9	1	6	4	8	9	12	-7	-3	-8	16 $\frac{1}{2}$	8	15
10	6	6	3	4	1	2	+2	+5	+1	3 $\frac{1}{2}$	14 $\frac{1}{2}$	5
11	4	0	2	8	9	12	-4	-9	-10	10	21	18 $\frac{1}{2}$
12	7	5	0	0	0	0	+7	+5	0	16 $\frac{1}{2}$	14 $\frac{1}{2}$	
13	0	2	12	8	9	12	-8	-7	0	19 $\frac{1}{2}$	19 $\frac{1}{2}$	2 $\frac{1}{2}$
14	2	3	12	6	8	12	-4	-5	0	10	14 $\frac{1}{2}$	2 $\frac{1}{2}$
15	0	0	12	7	4	6	-7	-4	+6	16 $\frac{1}{2}$	11	12 $\frac{1}{2}$
16	0	0	2	3	6	11	-3	-6	-9	6	17 $\frac{1}{2}$	16 $\frac{1}{2}$
17	7	6	12	8	6	12	-1	0	0	1 $\frac{1}{2}$	2 $\frac{1}{2}$	2 $\frac{1}{2}$
18	0	0	12	0	0	12	0	0	0	1 $\frac{1}{2}$	2 $\frac{1}{2}$	2 $\frac{1}{2}$
19	8	5	4	2	4	10	+6	+1	-6	13 $\frac{1}{2}$	5	12 $\frac{1}{2}$
20	0	0	2	0	0	11	0	0	-9	1 $\frac{1}{2}$	2 $\frac{1}{2}$	16 $\frac{1}{2}$
21	6	8	9	4	5	5	+2	+3	+4	3 $\frac{1}{2}$	8	8 $\frac{1}{2}$
22	0	0	2	1	0	7	-1	0	-5	1 $\frac{1}{2}$		10 $\frac{1}{2}$

Errors of Omission

Sub- ject	<u>GR group</u>			<u>PR group</u>			<u>Differences</u>			<u>Ranks</u>		
	Ah	A	F	Ah	A	F	Ah	A	F	Ah	A	F
1	0	3	0	2	2	2	-2	+1	-2	9 $\frac{1}{2}$	10 $\frac{1}{2}$	8 $\frac{1}{2}$
2	11	9	3	8	11	2	+3	-2	+1	12	14	3 $\frac{1}{2}$
3	10	11	1	9	11	5	+1	0	-4	7	4 $\frac{1}{2}$	15 $\frac{1}{2}$
4	2	3	0	12	11	5	-10	-8	-5	19 $\frac{1}{2}$	18 $\frac{1}{2}$	18
5	12	10	6	12	12	8	0	-2	-2	3 $\frac{1}{2}$	14	8 $\frac{1}{2}$
6	7	6	3	7	6	1	0	0	+2	3 $\frac{1}{2}$	4 $\frac{1}{2}$	8 $\frac{1}{2}$
7	0	0	0	0	0	0	0	0	0	3 $\frac{1}{2}$	4 $\frac{1}{2}$	
8	8	6	2	10	11	0	-2	-5	+2	9 $\frac{1}{2}$	17	8 $\frac{1}{2}$
9	12	11	5	8	11	8	+4	0	-3	13 $\frac{1}{2}$	4 $\frac{1}{2}$	13
10	2	3	0	10	6	4	-8	-3	-4	18	16	15 $\frac{1}{2}$
11	12	11	0	12	11	8	0	0	-8	3 $\frac{1}{2}$	4 $\frac{1}{2}$	20 $\frac{1}{2}$
12	8	11	1	12	11	1	-4	0	0	13 $\frac{1}{2}$	4 $\frac{1}{2}$	1 $\frac{1}{2}$
13	7	11	0	12	9	2	-5	+2	-2	15 $\frac{1}{2}$	14	8 $\frac{1}{2}$
14	10	11	8	4	2	4	+6	+9	+4	17	16 $\frac{1}{2}$	15 $\frac{1}{2}$
15	11	11	8	6	10	7	+5	+1	+1	15 $\frac{1}{2}$	6 $\frac{1}{2}$	3 $\frac{1}{2}$
16	4	0	2	2	1	4	+2	-1	-2	9 $\frac{1}{2}$	6 $\frac{1}{2}$	8 $\frac{1}{2}$
17	0	2	0	10	10	4	-10	-8	-4	19 $\frac{1}{2}$	14 $\frac{1}{2}$	15 $\frac{1}{2}$
18	12	11	0	12	11	8	0	0	-8	3 $\frac{1}{2}$	4 $\frac{1}{2}$	20 $\frac{1}{2}$
19	1	0	0	12	9	7	-11	-9	-7	21	16 $\frac{1}{2}$	19
20	12	11	2	12	11	0	0	0	+2	3 $\frac{1}{2}$	4 $\frac{1}{2}$	8 $\frac{1}{2}$
21	9	10	3	11	11	3	-2	-1	0	9 $\frac{1}{2}$	6 $\frac{1}{2}$	1 $\frac{1}{2}$
22	12	11	6	12	11	4	0	0	+2			8 $\frac{1}{2}$

(j) Number of different patterns per statement-type. A 'pattern' is the set of four pluses and/or minuses which characterises a subject's response to a statement-type on any one trial, as in (i) above.

Sub- ject	<u>GR group</u>						<u>PR group</u>					
	Statement Type					Total	Statement Type					Total
Ah	Ad	A	E	F	Ah		Ad	A	E	F		
1	2	2	3	2	3	12	2	1	2	2	3	10
2	3	2	4	5	2	16	3	2	3	2	3	13
3	3	2	3	2	4	14	2	4	3	3	4	16
4	2	1	2	2	1	8	1	2	2	2	4	11
5	3	3	2	3	3	14	1	1	1	1	1	5
6	2	3	2	3	3	13	3	3	3	2	3	14
7	1	1	2	2	2	8	1	1	1	2	2	7
8	2	2	4	2	4	14	2	1	1	2	2	8
9	2	2	2	2	3	11	2	1	1	2	2	8
10	3	2	3	1	2	11	3	3	4	3	4	17
11	2	1	1	3	3	10	1	2	1	3	2	9
12	4	2	2	3	3	14	1	3	1	2	4	11
13	3	1	2	2	1	9	1	1	1	2	3	8
14	3	2	2	2	1	10	3	3	4	3	3	16
15	1	1	1	3	2	8	2	2	2	2	4	12
16	2	1	2	3	2	10	3	1	3	4	4	15
17	1	1	3	2	2	9	2	2	2	1	3	10
18	1	4	1	1	3	10	1	1	1	1	1	5
19	2	1	2	1	3	9	2	5	3	3	3	16
20	1	5	1	2	5	14	1	2	1	1	3	8
21	2	2	2	3	3	12	2	3	1	2	3	11
22	1	1	1	1	2	6	2	1	2	3	2	10
	46	42	50	50	54	242	41	45	43	48	63	240

(k) 'Agreement scores' for the A-Ah and A-F pairs of statements and differences within groups between pairs and between groups for each pair. Since the former are significant on the Sign Test, difference ranks are presented only for the latter.

	<u>GR group</u>		<u>PR group</u>		<u>Differences</u>				<u>Ranks</u>	
	A-Ah	A-F	A-Ah	A-F	Within		Between		A-Ah	A-F
					GR	PR	A-Ah	A-F		
1	17	11	19	5	+6	+14	-2	+6	9	11
2	9	12	17	6	-3	+11	-8	+6	20 $\frac{1}{2}$	11
3	14	8	16	15	+6	+1	-2	-7	9	13
4	20	2	19	17	+18	+2	+1	-15	4	21
5	18	8	20	20	+10	0	-2	-12	9	18 $\frac{1}{2}$
6	20	20	16	14	0	+2	+4	+6	4	11
7	19	11	20	2	+8	+18	-1	+9	4	16
8	14	6	17	14	+8	+3	-3	-8	14	14 $\frac{1}{2}$
9	18	9	18	19	+9	-1	0	-10	1 $\frac{1}{2}$	17
10	18	7	10	6	+11	+4	+8	+1	20 $\frac{1}{2}$	1
11	18	6	20	19	+12	+1	-2	-13	9	20
12	16	9	20	7	+7	+13	-4	+2	7	2 $\frac{1}{2}$
13	16	11	19	15	+5	+4	-3	-4	14	6
14	18	17	15	9	+1	+6	+3	+8	14	14 $\frac{1}{2}$
15	20	18	16	14	+2	+2	+4	+4	17	6
16	18	5	13	10	+13	+3	+5	-5	19	9
17	16	9	14	13	+7	+1	+2	-4	9	6
18	20	10	20	10	+10	+10	0	0	1 $\frac{1}{2}$	
19	15	4	13	8	+11	+5	+2	-4	9	6
20	20	9	20	13	+11	+7	0	-4		6
21	19	8	20	20	+11	0	-1	-12	4	17
22	19	10	17	12	+9	+5	+2	-2	9	2 $\frac{1}{2}$

(1) 'Matching responses' in the four statement-types where these are not identical with the correct response. As explained in the text, the straightforward identification of such responses as ++-- in Ah and A, ---- in Ad and ---+ in F is complicated by the fact that the minus can represent an incomplete response. Some subjects have therefore ---- responses in Ad which are not in fact matching responses and have not, of course, been included in the following tabulation. The subjects concerned are 5, 6 and 21 in the GR group and 10, 14 and 19 in the PR group.

Sub- ject	GR group					PR group					Total Diff.	Rank
	<u>Statement Type</u>					<u>Statement Type</u>						
	Ah	Ad	A	F	Total	Ah	Ad	A	F	Total		
1	0	0	0	0	0	0	0	1	0	1	-1	4 $\frac{1}{2}$
2	1	0	1	0	2	3	0	3	1	7	-5	10
3	1	0	2	0	3	0	0	2	1	3	0	1 $\frac{1}{2}$
4	0	0	0	0	0	4	0	4	2	10	-10	18 $\frac{1}{2}$
5	2	0	4	3	9	5	0	5	5	15	-6	12
6	0	0	0	0	0	2	1	3	0	6	-6	12
7	0	0	0	0	0	0	0	0	0	0	0	1 $\frac{1}{2}$
8	4	1	1	1	7	0	0	0	0	0	+7	14 $\frac{1}{2}$
9	1	0	3	2	6	4	5	5	4	18	-12	21
10	1	0	2	0	3	0	0	1	0	1	+2	7
11	2	0	0	0	2	5	2	5	4	16	-14	22
12	2	1	3	0	6	0	2	0	0	2	+4	8
13	0	0	1	0	1	5	0	5	1	11	-10	18 $\frac{1}{2}$
14	1	0	2	5	8	0	0	1	0	1	+7	14 $\frac{1}{2}$
15	0	0	0	0	0	4	2	3	2	11	-11	20
16	0	0	0	0	0	0	0	0	1	1	-1	4 $\frac{1}{2}$
17	0	0	0	0	0	4	0	4	1	9	-9	17
18	0	0	0	0	0	0	0	0	5	5	-5	10
19	0	0	0	0	0	0	1	0	0	1	-1	4 $\frac{1}{2}$
20	0	1	0	0	1	0	0	0	0	0	+1	4 $\frac{1}{2}$
21	4	0	4	1	9	1	0	0	0	1	+8	16
22	0	0	0	0	0	0	5	0	0	5	-5	10
					57					124		

APPENDIX G

RAW DATA RELATING TO THE 4TS TASK

(a) Scores on the two half-sets (L and R) on the four selections of Ah, Ad and F and on the first ten selections of A. In the 4TS experiment the GR member of pair 7 was unable to take part. Accordingly, the pairs in Appendix G are displaced one row upwards, as compared with Appendices F and H, from 7 onwards. In the F type of statement, moreover, the PR member of (the 4TS) pair 9 could not be persuaded to complete the task. There are, therefore, no results for this pair in the relevant places in the tables which follow. In the intergroup differences scores on the two half-sets, separately and together, are taken into account.

SUB- JECT	<u>GR group</u>								<u>PR group</u>								
	<u>Statement Type</u>								<u>Statement Type</u>								
	Ah		Ad		A		F		Ah		Ad		A		F		
L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	H
1	8	8	8	8	20	20	8	8	8	3	8	8	20	18	8	8	
2	6	0	6	6	14	7	3	2	8	0	8	8	19	2	4	5	
3	8	2	8	8	18	16	8	8	8	2	8	8	20	18	4	8	
4	8	8	8	8	20	20	8	8	8	0	8	8	20	6	4	8	
5	8	4	5	5	16	6	5	2	8	0	8	8	19	2	6	8	
6	8	3	8	8	20	17	4	8	8	0	6	8	16	0	0	5	
7	5	2	8	8	20	13	4	8	8	4	8	8	20	10	4	4	
8	8	4	8	8	20	16	8	8	8	0	0	0	20	0	0	4	
9	8	2	8	8	20	14			8	0	0	1	16	6			
10	8	8	8	8	20	17	5	7	8	4	8	8	20	6	0	8	
11	8	2	5	5	20	14	8	8	8	2	8	8	20	12	8	8	
12	8	3	8	8	20	12	6	8	8	3	8	8	20	11	6	6	
13	8	3	8	8	20	8	3	8	4	0	6	6	18	4	4	0	
14	4	4	7	7	17	8	4	6	8	0	8	8	18	8	8	8	
15	8	8	8	8	20	20	8	8	6	4	8	8	20	10	2	4	
16	8	6	8	8	20	20	8	8	8	0	7	7	18	8	6	5	
17	8	8	8	8	20	20	8	8	8	3	8	8	19	8	4	6	
18	8	8	8	8	20	20	8	8	4	2	2	5	16	3	4	2	
19	8	8	8	8	18	20	8	8	8	3	8	8	20	20	7	8	
20	8	0	8	8	18	4	8	8	8	0	7	8	18	7	1	7	
21	8	4	8	8	20	14	8	8	2	4	8	8	19	2	0	4	

Intergroup Differences

	Ah			Ad			A			F		
	L+R	L	R	L+R	L	R	L+R	L	R	L+R	L	R
1	+5	0	+5	0	0	0	+2	0	+2	0	0	0
2	-2	-2	0	-4	-2	-2	0	-5	+5	-4	-1	-3
3	0	0	0	0	0	0	-4	-2	-2	+4	+4	0
4	+8	0	+8	0	0	0	+14	0	+14	+4	+4	0
5	+4	0	+4	-6	-3	-3	+1	-3	+4	-7	-1	-6
6	+3	0	+3	+2	+2	0	+21	+4	+17	+7	+4	+3
7	-5	-3	-2	0	0	0	+3	0	+3	+4	0	+4
8	+4	0	+4	+16	+8	+8	+16	0	+16	+12	+8	+4
9	+2	0	+2	+15	+8	+7	+12	+4	+8			
10	+4	0	+4	0	0	0	+11	0	+11	+4	+5	-1
11	0	0	0	-6	-3	-3	+2	0	+2	0	0	0
12	0	0	0	0	0	0	+1	0	+1	+2	0	+2
13	+7	+4	+3	+4	+2	+2	+6	+2	+4	+7	-1	+8
14	0	-4	+4	-2	-1	-1	-1	-1	0	-6	-4	-2
15	+6	+2	+4	0	0	0	+10	0	+10	+10	+6	+4
16	+6	0	+6	+2	+1	+1	+14	+2	+12	+5	+2	+3
17	+5	0	+5	0	0	0	+13	+1	+12	+6	+4	+2
18	+10	+4	+6	+9	+6	+3	+21	+4	+17	+10	+4	+6
19	+5	0	+5	0	0	0	-2	-2	0	+1	+1	0
20	0	0	0	+1	+1	0	-3	0	-3	+8	+7	+1
21	+6	+6	0	0	0	0	+13	+1	+12	+12	+8	+4

Difference Ranks

1	12 $\frac{1}{2}$	7 $\frac{1}{2}$	17	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	5	5	1 $\frac{1}{2}$	2 $\frac{1}{2}$		
2	5 $\frac{1}{2}$	15 $\frac{1}{2}$	3 $\frac{1}{2}$	15 $\frac{1}{2}$	15	15 $\frac{1}{2}$		20	11	7	6 $\frac{1}{2}$	11
3		7 $\frac{1}{2}$	3 $\frac{1}{2}$	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	9	13 $\frac{1}{2}$	5	7	12 $\frac{1}{2}$	2 $\frac{1}{2}$
4	19	7 $\frac{1}{2}$	21	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	16 $\frac{1}{2}$	4 $\frac{1}{2}$	18	7	12 $\frac{1}{2}$	2 $\frac{1}{2}$
5	9	7 $\frac{1}{2}$	13	17 $\frac{1}{2}$	17 $\frac{1}{2}$	18	2	16	9 $\frac{1}{2}$	14	6 $\frac{1}{2}$	17 $\frac{1}{2}$
6	7	7 $\frac{1}{2}$	9 $\frac{1}{2}$	13	15	6 $\frac{1}{2}$	19 $\frac{1}{2}$	18	19 $\frac{1}{2}$	14	12 $\frac{1}{2}$	11
7	12 $\frac{1}{2}$	17	7 $\frac{1}{2}$	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	7 $\frac{1}{2}$	4 $\frac{1}{2}$	7 $\frac{1}{2}$	7 $\frac{1}{2}$	2 $\frac{1}{2}$	14 $\frac{1}{2}$
8	9	7 $\frac{1}{2}$	13	21	20 $\frac{1}{2}$	21	18	4 $\frac{1}{2}$	19	19 $\frac{1}{2}$	19 $\frac{1}{2}$	14 $\frac{1}{2}$
9	5 $\frac{1}{2}$	7 $\frac{1}{2}$	7 $\frac{1}{2}$	20	20 $\frac{1}{2}$	20	13	18	12			
10	9	7 $\frac{1}{2}$	13	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	12	4 $\frac{1}{2}$	14	7	16	5 $\frac{1}{2}$
11	2 $\frac{1}{2}$	7 $\frac{1}{2}$	3 $\frac{1}{2}$	17 $\frac{1}{2}$	17 $\frac{1}{2}$	18	5	4 $\frac{1}{2}$	5	1 $\frac{1}{2}$	2 $\frac{1}{2}$	2 $\frac{1}{2}$
12	2 $\frac{1}{2}$	7 $\frac{1}{2}$	3 $\frac{1}{2}$	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	2	4 $\frac{1}{2}$	3	4	2 $\frac{1}{2}$	8
13	18	19	9 $\frac{1}{2}$	15 $\frac{1}{2}$	15	15 $\frac{1}{2}$	10	13 $\frac{1}{2}$	9 $\frac{1}{2}$	14	6 $\frac{1}{2}$	19
14	2 $\frac{1}{2}$	19	13	13	12	13 $\frac{1}{2}$	2	10	1 $\frac{1}{2}$	11 $\frac{1}{2}$	12 $\frac{1}{2}$	8
15	16	15 $\frac{1}{2}$	13	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	11	4 $\frac{1}{2}$	13	17 $\frac{1}{2}$	17	14 $\frac{1}{2}$
16	16	7 $\frac{1}{2}$	19 $\frac{1}{2}$	13	12	13 $\frac{1}{2}$	16 $\frac{1}{2}$	13 $\frac{1}{2}$	16	10	9	11
17	12 $\frac{1}{2}$	7 $\frac{1}{2}$	17	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	14 $\frac{1}{2}$	10	16	11 $\frac{1}{2}$	12 $\frac{1}{2}$	8
18	20	19	19 $\frac{1}{2}$	19	19	18	19 $\frac{1}{2}$	18	20	17 $\frac{1}{2}$	12 $\frac{1}{2}$	17 $\frac{1}{2}$
19	12 $\frac{1}{2}$	7 $\frac{1}{2}$	17	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	5	13 $\frac{1}{2}$	1 $\frac{1}{2}$	3	6 $\frac{1}{2}$	2 $\frac{1}{2}$
20	2 $\frac{1}{2}$	7 $\frac{1}{2}$	3 $\frac{1}{2}$	11	12	6 $\frac{1}{2}$	7 $\frac{1}{2}$	4 $\frac{1}{2}$	7 $\frac{1}{2}$	16	18	5 $\frac{1}{2}$
21	16	21	3 $\frac{1}{2}$	5 $\frac{1}{2}$	5 $\frac{1}{2}$	6 $\frac{1}{2}$	14 $\frac{1}{2}$	10	16	19 $\frac{1}{2}$	19 $\frac{1}{2}$	14 $\frac{1}{2}$

(b) Scores on the four selections of F, both half-sets together and left-hand half-set separately, intergroup differences and difference ranks.

	<u>GR group</u>								<u>PR group</u>							
	<u>L and R</u>				<u>L only</u>				<u>L and R</u>				<u>L only</u>			
	<u>S e l e c t i o n s</u>								<u>S e l e c t i o n s</u>							
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	4	4	4	4	2	2	2	2	4	4	4	4	2	2	2	2
2	0	0	2	3	0	0	2	1	2	2	2	3	1	1	1	1
3	4	4	4	4	2	2	2	2	3	3	3	3	1	1	1	1
4	4	4	4	4	2	2	2	2	3	3	3	3	1	1	1	1
5	1	2	2	2	1	1	2	1	3	4	3	4	1	1	2	2
6	2	2	4	4	0	0	2	2	1	1	1	2	0	0	0	0
7	2	2	4	4	0	0	2	2	2	2	2	2	1	1	1	1
8	4	4	4	4	2	2	2	2	1	1	0	2	0	0	0	0
10	3	3	3	3	1	1	2	1	2	2	2	2	0	0	0	0
11	4	4	4	4	2	2	2	2	4	4	4	4	2	2	2	2
12	3	3	4	4	1	1	2	2	3	2	3	3	2	2	1	1
13	2	3	3	3	0	1	1	1	1	1	1	1	1	1	1	1
14	2	2	3	3	1	1	1	1	4	4	4	4	2	2	2	2
15	4	4	4	4	2	2	2	2	1	2	2	2	0	0	1	1
16	4	4	4	4	2	2	2	2	2	3	3	4	1	1	2	2
17	4	4	4	4	2	2	2	2	3	3	2	2	1	1	1	1
18	4	4	4	4	2	2	2	2	2	0	1	1	2	0	0	0
19	4	4	4	4	2	2	2	2	4	4	4	4	2	2	2	2
20	4	4	4	4	2	2	2	2	2	2	2	2	1	0	0	0
21	4	4	4	4	2	2	2	2	1	1	1	1	0	0	0	0

	<u>Intergroup Differences</u>								<u>Difference Ranks</u>								
	L and R				L only				L and R				L only				
1	0	0	0	0	0	0	0	0		$2\frac{1}{2}$	$2\frac{1}{2}$			$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
2	-2	-2	0	0	-1	-1	+1	0	$13\frac{1}{2}$	$14\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	12	$11\frac{1}{2}$	10	$3\frac{1}{2}$	$3\frac{1}{2}$
3	+1	+1	+1	+1	+1	+1	+1	+1	$7\frac{1}{2}$	8	8	8	12	$11\frac{1}{2}$	10	11	11
4	+1	+1	+1	+1	+1	+1	+1	+1	$7\frac{1}{2}$	8	8	8	12	$11\frac{1}{2}$	10	11	11
5	-2	-2	-1	-2	0	0	0	-1	$13\frac{1}{2}$	$14\frac{1}{2}$	8	$14\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	11	11
6	+1	+1	+3	+2	0	0	+2	+2	$7\frac{1}{2}$	8	18	$14\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$17\frac{1}{2}$	18	18
7	0	0	+2	+2	-1	-1	+1	+1	$2\frac{1}{2}$	$2\frac{1}{2}$	$13\frac{1}{2}$	$14\frac{1}{2}$	12	$11\frac{1}{2}$	10	11	11
8	+3	+3	+4	+2	+2	+2	+2	+2	18	$18\frac{1}{2}$	20	$14\frac{1}{2}$	19	$18\frac{1}{2}$	$17\frac{1}{2}$	18	18
10	+1	+1	+1	+1	+1	+1	+2	+1	$7\frac{1}{2}$	8	8	8	12	$11\frac{1}{2}$	$17\frac{1}{2}$	11	11
11	0	0	0	0	0	0	0	0	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
12	0	+1	+1	+1	-1	-1	+1	+1	$2\frac{1}{2}$	8	8	8	12	$11\frac{1}{2}$	10	11	11
13	+1	+2	+2	+2	-1	0	0	0	$7\frac{1}{2}$	$14\frac{1}{2}$	$13\frac{1}{2}$	$14\frac{1}{2}$	12	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
14	-2	-2	-1	-1	-1	-1	-1	-1	$13\frac{1}{2}$	$14\frac{1}{2}$	8	8	12	$11\frac{1}{2}$	10	11	11
15	+3	+2	+2	+2	+2	+2	+1	+1	18	$14\frac{1}{2}$	$13\frac{1}{2}$	$14\frac{1}{2}$	19	$18\frac{1}{2}$	10	11	11
16	+2	+1	+1	0	+1	+1	0	0	$13\frac{1}{2}$	8	8	$2\frac{1}{2}$	12	$11\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
17	+1	+1	+2	+2	+1	+1	+1	+1	$7\frac{1}{2}$	8	$13\frac{1}{2}$	$14\frac{1}{2}$	12	$11\frac{1}{2}$	10	11	11
18	+2	+4	+3	+3	0	+2	+2	+2	$13\frac{1}{2}$	20	18	$19\frac{1}{2}$	$3\frac{1}{2}$	$18\frac{1}{2}$	$17\frac{1}{2}$	18	18
19	0	0	0	0	0	0	0	0	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
20	+2	+2	+2	+2	+1	+2	+2	+2	$13\frac{1}{2}$	$14\frac{1}{2}$	$13\frac{1}{2}$	$14\frac{1}{2}$	12	$18\frac{1}{2}$	$17\frac{1}{2}$	18	18
21	+3	+3	+3	+3	+2	+2	+2	+2	18	$18\frac{1}{2}$	18	$19\frac{1}{2}$	19	$18\frac{1}{2}$	$17\frac{1}{2}$	18	18

(c) 'Converse errors' in Ah, A and both statement-types together.

	<u>GR group</u>			<u>PR group</u>			<u>Differences</u>			<u>Difference Ranks</u>		
	Ah	A	Both	Ah	A	Both	Ah	A	Both	Ah	A	Both
1	0	0	0	0	2	2	0	-2	-2		9	7
2	1	2	3	2	5	7	-1	-3	-4	$7\frac{1}{2}$	12	$12\frac{1}{2}$
3	0	$1\frac{1}{2}$	$1\frac{1}{2}$	1	0	1	-1	$+1\frac{1}{2}$	$+\frac{1}{2}$	$7\frac{1}{2}$	7	3
4	0	0	0	2	4	6	-2	-4	-6	$15\frac{1}{2}$	$14\frac{1}{2}$	15
5	0	8	8	2	4	6	-2	+4	+2	$15\frac{1}{2}$	$14\frac{1}{2}$	7
6	0	0	0	2	2	4	-2	-2	-4	$15\frac{1}{2}$	9	$12\frac{1}{2}$
7	2	0	2	0	0	0	+2	0	+2	$15\frac{1}{2}$		7
8	0	0	0	2	10	12	-2	-10	-12	$15\frac{1}{2}$	20	20
9	1	2	3	2	8	10	-1	-6	-7	$7\frac{1}{2}$	17	$16\frac{1}{2}$
10	2	1	3	1	2	3	+1	-1	0	$7\frac{1}{2}$	$4\frac{1}{2}$	$1\frac{1}{2}$
11	2	1	3	0	0	0	+2	+1	+3	$15\frac{1}{2}$	$4\frac{1}{2}$	10
12	2	0	2	1	2	3	+1	-2	-1	$7\frac{1}{2}$	9	$4\frac{1}{2}$
13	0	0	0	2	7	9	-2	-7	-9	$15\frac{1}{2}$	$18\frac{1}{2}$	19
14	2	1	3	2	0	2	0	+1	+1	$2\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$
15	0	0	0	2	1	3	-2	-1	-3	$15\frac{1}{2}$	$4\frac{1}{2}$	10
16	0	0	0	2	3	5	-2	-3	-5	$15\frac{1}{2}$	12	14
17	0	0	0	0	0	0	0	0	0	$2\frac{1}{2}$	$1\frac{1}{2}$	$1\frac{1}{2}$
18	0	2	2	2	7	9	-2	-5	-7	$15\frac{1}{2}$	16	$16\frac{1}{2}$
19	0	0	0	0	0	0	0	0	0	$2\frac{1}{2}$	$1\frac{1}{2}$	$1\frac{1}{2}$
20	1	1	2	1	4	5	0	-3	-3	$2\frac{1}{2}$	12	10
21	0	0	0	1	7	8	-1	-7	-8	$7\frac{1}{2}$	$18\frac{1}{2}$	18

APPENDIX H

RAW DATA RELATING TO THE NEGATIVES TASK AND SCORES ON THE EPI

(a) Response times in tenths of a second on six successive trials on the four conditions, true affirmative (T+), true negative (T-), false affirmative (F+) and false negative (F-).

Sub- ject	<u>GR group</u>						<u>True Affirmative</u>						<u>PR group</u>					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	22	24	24	22	26	14	20	19	36	22	13	18	20	19	36	22	13	18
2	18	27	23	20	22	26	46	12	17	13	16	15	46	12	17	13	16	15
3	16	18	14	23	12	15	16	14	14	16	16	15	16	14	14	16	16	15
4	24	29	15	18	20	21	23	17	17	14	15	13	23	17	17	14	15	13
5	37	31	22	27	23	26	13	19	18	19	18	15	13	19	18	19	18	15
6	15	21	21	27	22	18	30	28	22	22	26	18	30	28	22	22	26	18
7	8	12	15	33	15	13	18	15	23	15	13	14	18	15	23	15	13	14
8	17	11	13	16	24	16	38	22	17	20	18	24	38	22	17	20	18	24
9	25	14	20	20	58	12	32	35	25	25	37	14	32	35	25	25	37	14
10	20	18	20	17	22	14	20	19	18	28	19	20	20	19	18	28	19	20
11	16	15	11	13	16	13	20	26	21	20	22	16	20	26	21	20	22	16
12	21	22	22	25	13	22	23	20	13	22	9	18	23	20	13	22	9	18
13	40	32	42	28	24	33	22	21	18	20	18	17	22	21	18	20	18	17
14	12	14	12	18	12	18	25	24	25	23	26	15	25	24	25	23	26	15
15	23	21	15	15	15	42	35	25	28	26	26	23	35	25	28	26	26	23
16	25	35	40	45	44	24	23	19	34	37	25	16	23	19	34	37	25	16
17	16	13	15	18	18	17	34	34	25	20	21	25	34	34	25	20	21	25
18	20	17	14	14	17	15	51	97	47	37	36	50	51	97	47	37	36	50
19	22	13	21	19	17	19	18	17	20	10	12	13	18	17	20	10	12	13
20	13	13	16	17	15	13	18	38	54	34	28	33	18	38	54	34	28	33
21	15	14	16	16	16	11	18	13	17	28	15	26	18	13	17	28	15	26
22	33	29	39	38	31	25	27	14	20	27	18	17	27	14	20	27	18	17

	<u>True Negative</u>																	
1	36	32	45	34	38	30	35	91	28	17	43	71	35	91	28	17	43	71
2	28	21	32	25	28	18	12	31	19	14	13	16	12	31	19	14	13	16
3	44	50	38	21	50	30	24	18	24	55	26	24	24	18	24	55	26	24
4	25	25	32	36	25	17	35	50	28	31	23	45	35	50	28	31	23	45
5	90	37	32	31	58	59	45	26	36	23	26	24	45	26	36	23	26	24
6	41	37	43	24	28	28	53	86	67	40	45	35	53	86	67	40	45	35
7	15	19	67	33	27	13	23	25	32	32	34	35	23	25	32	32	34	35
8	24	14	18	26	14	20	78	24	20	24	24	34	78	24	20	24	24	34
9	40	59	30	24	42	42	35	44	27	49	37	42	35	44	27	49	37	42
10	20	24	20	17	41	20	33	36	34	34	31	34	33	36	34	34	31	34
11	15	15	76	20	21	29	45	59	47	25	45	30	45	59	47	25	45	30
12	30	36	81	47	55	34	67	30	34	25	48	22	67	30	34	25	48	22
13	124	111	73	55	37	28	57	135	29	25	30	32	57	135	29	25	30	32
14	24	22	22	12	14	20	40	27	37	31	27	30	40	27	37	31	27	30
15	21	23	17	24	37	26	42	31	46	43	37	29	42	31	46	43	37	29
16	40	42	48	34	60	35	87	40	29	40	20	32	87	40	29	40	20	32
17	33	30	42	20	48	39	38	60	41	46	50	52	38	60	41	46	50	52
18	78	54	15	22	19	24	76	75	55	51	48	73	76	75	55	51	48	73
19	35	23	27	28	46	52	68	57	36	42	22	28	68	57	36	42	22	28
20	22	17	22	28	15	17	45	22	45	34	30	21	45	22	45	34	30	21
21	24	68	22	34	27	19	96	35	43	52	79	45	96	35	43	52	79	45
22	52	44	73	50	67	43	40	54	33	50	33	181	40	54	33	50	33	181

	<u>GR_group</u>						<u>False Affirmative</u>						<u>PR_group</u>					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	26	25	32	31	38	25	18	17	23	26	65	20						
2	23	22	33	21	33	28	15	21	27	16	18	16						
3	15	22	41	24	18	24	17	16	16	21	62	16						
4	38	20	21	18	26	18	19	27	37	15	17	16						
5	28	68	46	31	22	27	17	32	32	27	16	20						
6	19	23	29	27	29	18	28	32	26	25	30	26						
7	11	18	21	17	21	14	23	39	26	28	22	44						
8	20	18	15	11	13	18	54	42	120	32	43	48						
9	43	40	47	38	24	46	32	54	32	36	34	68						
10	18	20	18	22	17	21	32	24	23	22	32	27						
11	15	15	16	19	15	16	15	24	33	18	30	29						
12	25	38	26	33	23	26	18	21	25	17	27	15						
13	20	33	28	30	35	28	22	27	34	22	28	25						
14	16	17	17	13	14	47	30	50	23	39	39	65						
15	28	22	15	18	19	28	29	28	37	29	29	26						
16	25	33	36	38	42	33	45	37	29	32	21	23						
17	17	17	17	22	13	23	54	44	35	21	27	23						
18	20	17	17	19	20	15	63	54	55	70	52	45						
19	27	19	22	45	23	30	38	16	26	16	15	15						
20	13	18	18	17	19	12	18	23	35	19	44	20						
21	19	23	14	17	19	22	37	51	45	30	60	41						
22	35	60	45	37	35	26	21	31	23	30	30	21						

	<u>False Negative</u>											
	1	2	3	4	5	6	1	2	3	4	5	6
1	40	102	57	74	50	36	37	65	34	44	28	46
2	37	22	33	52	63	52	38	15	57	64	18	37
3	241	47	32	134	45	30	52	24	22	46	40	33
4	74	74	160	39	125	15	95	257	23	17	23	75
5	369	62	62	199	159	87	25	20	27	27	32	70
6	39	24	27	22	21	25	40	81	52	35	106	358
7	44	18	181	121	30	27	50	28	47	43	36	36
8	33	58	44	15	29	22	62	33	34	28	60	25
9	50	377	74	78	62	59	48	51	46	239	82	35
10	26	14	20	17	21	21	115	32	35	45	32	34
11	43	16	17	20	19	23	32	41	29	28	34	42
12	70	46	33	42	38	32	22	37	47	50	60	25
13	54	90	120	66	72	114	103	66	48	115	44	29
14	48	15	20	18	29	15	45	45	56	62	47	67
15	52	26	25	25	82	68	60	45	77	56	40	66
16	56	62	205	41	40	50	91	137	46	92	67	28
17	24	25	17	44	41	100	134	65	63	37	88	36
18	40	65	33	20	15	21	66	130	100	99	42	63
19	22	69	47	29	35	40	96	39	157	128	122	223
20	20	22	21	24	24	15	20	15	23	57	23	42
21	29	24	18	50	388	41	49	38	98	71	124	72
22	60	64	59	78	82	49	304	287	55	139	35	35

(b) Total response times over six trials on the four conditions separately and on the negative conditions (T- and F-) together and the conditions involving a single negative component (T- and F+) together.

	<u>GR group</u>						<u>PR group</u>					
	T+	T-	F+	F-	T-/F-	T-/F+	T+	T-	F+	F-	T-/F-	T-/F+
1	132	215	177	359	574	392	123	197	149	180	377	346
2	199	428	174	516	944	602	125	251	149	206	457	400
3	87	121	97	126	247	218	90	253	126	765	1018	379
4	86	114	124	145	259	238	102	180	144	201	381	324
5	97	116	95	201	317	211	116	308	158	405	713	466
6	111	211	166	242	453	377	159	287	204	423	710	491
7	136	152	160	259	411	312	105	226	123	241	467	349
8	149	237	238	700	937	475	91	171	148	217	388	319
9	84	176	96	138	314	272	117	391	156	855	1246	547
10	98	233	144	529	762	377	139	204	339	242	446	543
11	88	194	114	550	744	308	138	192	246	322	514	438
12	127	160	141	487	647	301	99	212	131	498	710	343
13	131	148	130	278	426	278	98	181	182	240	421	363
14	97	212	109	251	463	321	124	202	160	293	495	362
15	96	174	102	421	595	276	119	105	113	229	334	218
16	124	201	145	158	359	346	205	350	264	452	802	614
17	166	307	222	938	1245	529	132	285	169	250	535	454
18	111	142	116	119	261	258	168	234	256	501	735	490
19	195	329	238	392	721	567	318	378	339	500	878	717
20	213	259	207	454	713	466	146	326	167	672	998	493
21	125	283	171	261	544	454	154	248	187	461	709	435
22	97	212	108	194	406	320	163	228	178	344	572	406

	<u>Intergroup Differences</u>						<u>Difference Ranks</u>					
	T+	T-	F+	F-	T-/F-	T-/F+	T+	T-	F+	F-	T-/F-	T-/F+
1	+9	+18	+28	+179	+197	+46	2	4	6	8	9	6
2	+74	+177	+25	+310	+487	+202	20	20	5	17	18	18
3	-3	-132	-29	-639	-771	-161	1	18	7	20	21	16
4	-16	-66	-20	-56	-121	-86	3	11½	4	5	5	10½
5	-19	-192	-63	-204	-396	-255	4	22	15	12	15	20
6	-48	-76	-38	-181	-257	-114	14	16	9	9	11	12
7	+31	-74	+37	+18	-56	-37	9	15	8	2	3	3
8	+58	+66	+90	+483	+549	+156	17	11½	17	19	19	15
9	+33	-215	-60	-717	-932	-275	10½	21	14	22	22	22
10	-41	+29	-195	+287	+316	-166	13	6	22	15	14	17
11	-50	+2	-132	+228	+230	-130	15	1	20	14	10	13
12	+28	-52	+10	-11	-63	-42	7	10	1	1	4	5
13	+33	-33	-52	+38	+5	-85	10½	7	12	3	1	9
14	-27	+10	-51	-42	-32	-41	6	2	11	4	2	4
15	-23	+69	-11	+192	+259	+58	5	14	2	10	12	7
16	-81	-149	-119	-294	-443	-258	21	19	19	16	16	21
17	+34	+22	+53	+688	+710	+75	12	5	13	21	20	8
18	-57	-92	-140	-382	-474	-232	16	17	21	18	17	19
19	-123	-49	-101	-108	-157	-150	22	9	18	6	6	14
20	+67	-67	+40	-218	-285	-27	19	13	10	13	13	2
21	-29	+35	-16	-200	-165	+19	8	8	3	11	7	1
22	-66	-16	-70	-150	-166	-86	18	3	16	7	8	10½

(c) Extraversion (E), Neuroticism (N) and Lie (L) scale scores on the A and B Forms, separately and together, of the Eysenck Personality Inventory. There are no scores for pair 9 on the B Form and the inter-group comparisons discussed in the text, as well as the rank correlations between the four conditions of the negatives task and E and N, are based on the remaining 21 pairs.

	<u>GR group</u>									<u>PR group</u>								
	Ea	Eb	E	Na	Nb	N	La	Lb	L	Ea	Eb	E	Na	Nb	N	La	Lb	L
1	9	13	22	19	12	31	2	0	2	9	9	18	18	19	37	3	0	3
2	16	10	26	14	17	31	0	1	1	4	10	14	12	14	26	4	0	4
3	13	10	23	12	8	20	0	0	0	13	16	29	9	5	14	0	0	0
4	10	8	18	14	14	28	0	0	0	12	14	26	9	12	21	2	0	2
5	8	11	19	11	10	21	4	1	5	8	10	18	18	15	33	3	0	3
6	12	10	22	11	13	24	1	0	1	9	12	21	13	16	29	2	1	3
7	6	11	17	16	16	32	3	1	0	10	17	27	12	12	24	3	0	3
8	14	12	26	13	21	34	1	0	0	6	9	15	13	20	33	4	0	4
9	11			5			3			8			6			4		
10	14	14	28	11	12	23	1	0	0	9	11	20	12	15	27	3	1	4
11	5	11	16	12	7	19	4	0	0	12	11	23	14	16	30	0	0	0
12	12	15	27	8	21	29	0	0	0	14	13	27	15	18	33	3	0	3
13	11	15	26	18	18	36	2	0	0	13	14	27	9	12	21	3	0	3
14	8	9	17	11	14	25	2	0	0	11	13	24	12	15	27	0	1	1
15	4	11	15	12	15	27	1	1	2	9	12	21	19	19	38	3	1	4
16	10	9	19	12	17	29	4	0	0	16	12	28	16	21	37	2	1	3
17	10	16	26	9	8	17	3	0	0	9	14	23	11	10	21	3	0	3
18	13	16	29	11	13	24	3	0	0	3	9	12	11	18	29	4	0	4
19	15	18	33	9	4	13	2	0	0	9	10	19	17	16	33	0	0	0
20	15	17	32	4	11	15	0	0	0	6	13	19	10	7	17	4	1	5
21	11	13	24	10	6	16	4	0	0	18	18	36	17	16	33	5	2	7
22	17	19	36	13	10	23	2	0	0	3	8	11	14	16	30	1	0	1

*233 268 501 250 267 517 39 4 43 203 255 458 281 312 593 52 8 60

* The sums for Form A do not include the scores for the members of pair 9.

(d) The ranks on which the Spearman's rho correlations between E and N and the four 'negatives' conditions were based. The groups are given in the usual order (GR group first) but with the members of the ninth pair omitted since the GR member did not do Form B of the EPI. Scores and response times are ranked from highest to lowest and from longest to shortest, respectively.

Sub- ject	E	N	T+	T-	F+	F-
1G	23 $\frac{1}{2}$	12 $\frac{1}{2}$	15 $\frac{1}{2}$	20	14	20
2G	15	12 $\frac{1}{2}$	4	1	15	8
3G	21	35	40	37	40	41
4G	33	20	41	41	33	39
5G	29 $\frac{1}{2}$	32 $\frac{1}{2}$	34	40	42	34 $\frac{1}{2}$
6G	23 $\frac{1}{2}$	27	27 $\frac{1}{2}$	24	19	28 $\frac{1}{2}$
7G	35 $\frac{1}{2}$	11	11	16	6 $\frac{1}{2}$	4
8G	15	5	42	33	41	40
9G	7 $\frac{1}{2}$	29 $\frac{1}{2}$	31 $\frac{1}{2}$	18	27 $\frac{1}{2}$	7
10G	37	36	39	29	35	6
11G	10 $\frac{1}{2}$	17 $\frac{1}{2}$	18	36	29	12
12G	15	4	17	38	31	24
13G	35 $\frac{1}{2}$	25	34	22	37	26
14G	38 $\frac{1}{2}$	22	36	34	39	17
15G	29 $\frac{1}{2}$	17 $\frac{1}{2}$	21 $\frac{1}{2}$	27	26	38
16G	15	37 $\frac{1}{2}$	7	8	8	1
17G	5 $\frac{1}{2}$	27	27 $\frac{1}{2}$	39	34	42
18G	3	42	5	5	6 $\frac{1}{2}$	19
19G	4	40	2	12	9	14
20G	18 $\frac{1}{2}$	39	19 $\frac{1}{2}$	11	16	25
21G	1 $\frac{1}{2}$	29 $\frac{1}{2}$	34	22	38	36
1P	33	2 $\frac{1}{2}$	23	28	23 $\frac{1}{2}$	37
2P	40	24	19 $\frac{1}{2}$	14	23 $\frac{1}{2}$	33
3P	5 $\frac{1}{2}$	41	38	13	32	3
4P	15	32 $\frac{1}{2}$	29	32	27 $\frac{1}{2}$	34 $\frac{1}{2}$
5P	33	8	26	7	21	18
6P	25 $\frac{1}{2}$	17 $\frac{1}{2}$	9	9	10	16
7P	10 $\frac{1}{2}$	27	37	35	25	32
8P	38 $\frac{1}{2}$	8	25	2	22	2
9P	27	22	13	25	1 $\frac{1}{2}$	28 $\frac{1}{2}$
10P	21	14 $\frac{1}{2}$	14	30	5	22
11P	10 $\frac{1}{2}$	8	30	22	30	11
12P	10 $\frac{1}{2}$	32 $\frac{1}{2}$	31 $\frac{1}{2}$	31	12	30
13P	18 $\frac{1}{2}$	22	21 $\frac{1}{2}$	26	20	23
14P	25 $\frac{1}{2}$	1	24	42	36	31
15P	7 $\frac{1}{2}$	2 $\frac{1}{2}$	3	4	3	15
16P	21	32 $\frac{1}{2}$	15 $\frac{1}{2}$	10	17	27
17P	41	17 $\frac{1}{2}$	6	17	4	9
18P	29 $\frac{1}{2}$	8	1	3	1 $\frac{1}{2}$	10
19P	29 $\frac{1}{2}$	37 $\frac{1}{2}$	12	6	18	5
20P	1 $\frac{1}{2}$	8	10	15	11	13
21P	42	14 $\frac{1}{2}$	8	19	13	21

BIBLIOGRAPHY

- Anstey, E. (1966), Psychological Tests. London: Nelson.
- Backhouse, J.K. (1967), The use of Valentine's Reasoning Tests in an investigation into transfer from mathematics to reasoning. Brit. J. educ. Psychol., 37, 121-3.
- Black, M. (1946), Critical Thinking. New York: Prentice-Hall.
- Bruner, J.S., Goodnow, J.J., and Austin, G.A. (1956), A Study of Thinking. New York: Wiley.
- Bruner, J.S., Olver, R.R. et al., (1966), Studies in Cognitive Growth. New York: Wiley.
- Butcher, H.J. (1968), Human Intelligence. London: Methuen.
- Box, G.E.P. (1953), Non-normality and tests on variances. Biometrika, 40, 318-35.
- Burt, C. (1919), The development of reasoning in school children. J. exp. Ped., 5, 68-77, 121-7.
- Burt, C. (1949), Mental and Scholastic Tests. Third Edition. London: Staples.
- Carney, J.D., and Scheer, R.K. (1965), Fundamentals of Logic. New York: Macmillan.
- Chapman, L.J. and Chapman, J.P. (1959), The atmosphere effect re-examined. J. exp. Psychol., 58, 220-6.
- Cohen, L.J. (1962), The Diversity of Meaning. London: Methuen.
- Diamond, S. (1959), Information and Error. New York: Basic Books.
- Donaldson, M.C. (1963), A Study of Children's Thinking. London: Tavistock.
- Edwards, A.L. (1968), Experimental Design in Psychological Research. New York: Holt, Rinehart and Winston.
- Eifermann, R.R. (1961), Negation: a linguistic variable. Acta Psychol., 18, 258-73.
- Elton, C.F. (1965), The effect of logic instruction on the Valentine Reasoning Test. Brit. J. educ. Psychol., 35, 339-41.
- Ennis, R.H. (1959), An evaluation of the Watson-Glaser 'Critical Thinking Appraisal'. J. educ. Res., 52, 155-8.
- Evans, E.G.S. (1964), Reasoning ability and personality differences among student teachers. Brit. J. educ. Psychol., 34, 305-14.

- Eysenck, H.J. (1958), Sense and Nonsense in Psychology. London: Penguin.
- Eysenck, H.J. (1962), Response set, authoritarianism and personality questionnaires. Brit. J. soc. clin. Psychol., 1, 20-4.
- Eysenck, H.J. (1964), Crime and Personality. London: Routledge and Kegan Paul.
- Eysenck, H.J. and Eysenck, S.B.G. (1964), Manual of the Eysenck Personality Inventory. London: University of London Press.
- Flavell, J.H. (1963), The Developmental Psychology of Jean Piaget. Princeton: Van Nostrand.
- Foss, B.M. (1966), editor, New Horizons in Psychology. London: Penguin.
- Friedman, N. (1967), The Social Nature of Psychological Research. New York: Basic Books.
- Gorden, R.L. (1953), The effect of attitude toward Russia on logical reasoning. J. soc. Psychol., 37, 103-11.
- Guilford, J.P. (1965), Fundamental Statistics in Psychology and Education. Fourth Edition. New York: McGraw-Hill.
- Guilford, J.P. (1967), The Nature of Human Intelligence. New York: McGraw-Hill.
- Hallworth, H.J. (1963), An analysis of C.W. Valentine's Reasoning Tests for Higher Levels of Intelligence. Brit. J. educ. Psychol., 33, 41-6.
- Harlow, H.F. (1949), The formation of learning sets. Psychol. Rev., 56, 51-65.
- Hays, W.L. (1963), Statistics for Psychologists. New York: Holt, Rinehart and Winston.
- Henle, M. (1962), On the relation between logic and thinking. Psychol. Rev., 69, 366-78.
- Henle, M. and Michael, M. (1956), The influence of attitudes on syllogistic reasoning. J. soc. Psychol., 44, 115-27.
- Hertzka, A.F. and Guilford, J.P. (1955), Logical Reasoning: a Manual of Instructions and Interpretations. Beverley Hills, Calif.: Sheridan Supply Company.
- Hughes, M.A.M. (1966), The Use of Negative Information in Concept Attainment. Unpublished University of London Ph.D. thesis.
- Hunt, J. McV. (1961), Intelligence and Experience. New York: Ronald.
- Hunter, I.M.L. (1957a), Note on an atmosphere effect in adult reasoning. Quart. J. exp. Psychol., 9, 175-6.

- Hunter, I.M.L. (1957b), The solving of three-term series problems. Brit. J. Psychol., 48, 286-98.
- Inhelder, B. and Piaget, J. (1958), The Growth of Logical Thinking. London: Routledge and Kegan Paul.
- Inhelder, B. and Piaget, J. (1964), The Early Growth of Logic in the Child. London: Routledge and Kegan Paul.
- Janis, I.L. and Frick, F. (1943), The relationship between attitudes towards conclusions and errors in judging the logical validity of syllogisms. J. exp. Psychol., 33, 73-7.
- Lefford, A. (1946), The influence of emotional subject-matter on logical reasoning. J. gen. Psychol., 34, 127-51.
- McNemar, Q. (1959), Psychological Statistics. Third Edition. New York: Wiley.
- Mill, J.S. (1843), A System of Logic. London.
- Nisbet, S.D. and Napier, B.L. (1970), Promise and Progress. Glasgow: Glasgow University Publication, NS 136.
- Nunnally, J.C. (1970), Introduction to Psychological Measurement. New York: McGraw-Hill.
- Osgood, C.E. (1957), The Measurement of Meaning. Urbana: Illinois University Press
- Peel, E.A. (1967), The Pupil's Thinking. Rev. Edition. London: Oldbourne Educational Books.
- Piaget, J. (1950), The Psychology of Intelligence. London: Routledge and Kegan Paul.
- Quine, W.V.O. (1952), Methods of Logic. London: Routledge and Kegan Paul.
- Raven, J.C. (1965), Advanced Progressive Matrices: Plan and Use of the Scale. London: H.K. Lewis.
- Richter, M.N. (1957), The theoretical interpretation of errors in syllogistic reasoning. J. Psychol., 43, 341-4.
- Ruby, L. (1968), The Art of Making Sense. Philadelphia: Lippincott.
- Sells, S.B. (1936), The atmosphere effect: an experimental study of reasoning. Arch. Psychol., 29, 3-72.
- Siegel, S. (1956), Non-parametric Statistics. New York: McGraw-Hill.
- Sinclair, W.A. (1951), The Traditional Formal Logic. Fifth Edition. London: Methuen.
- Stebbing, L.S. (1952), A Modern Elementary Logic. Fifth Edition. London: Methuen.
- Stebbing, L.S. (1959), Thinking to Some Purpose. London: Penguin.

- Strawson, P.F. (1952), Introduction to Logical Theory. London: Methuen.
- Terman, L.M. (1956), Manual of the Concept Mastery Test. New York: Psychological Corporation.
- Terman, L.M. and Oden, M.H. (1947), The Gifted Child Grows Up: Genetic Studies of Genius, IV. Stanford: Stanford University Press.
- Thistlethwaite, D. (1950), Attitude and structure as factors in the distortion of reasoning. J. abnorm. soc. Psychol., 45, 442-58.
- Thorndike, E.L. (1911), Animal Intelligence. New York: Macmillan.
- Thorndike, E.L. (1922), The effect of changed data on reasoning. J. exp. Psychol., 5, 33-8.
- Thouless, R.H. (1945), Straight and Crooked Thinking. London: Hodder and Stoughton.
- Valentine, C.W. (1957), Handbook of the Reasoning Tests for Higher Levels of Intelligence. Edinburgh: Oliver and Boyd.
- Valentine, C.W. (1961), The use of a new reasoning test for selection of university and training college students. Brit. J. educ. Psychol., 31, 227-31.
- Vernon, P.E. (1950), The Structure of Human Abilities. London: Methuen.
- Vernon, P.E. (1964), Personality Assessment. London: Methuen.
- Vinacke, W.E. (1957), The Psychology of Thinking. New York: McGraw-Hill.
- Wallace, I.G. (1965), Failure to Distinguish Valid from Invalid Arguments. Unpublished University of Glasgow Ed.B. thesis.
- Wason, P.C. (1959), The processing of positive and negative information. Quart. J. exp. Psychol., 11, 92-107
- Wason, P.C. (1961), Response to affirmative and negative binary statements. Brit. J. Psychol., 52, 133-42.
- Wason, P.C. (1966), Reasoning. (In Foss, 1966, above.)
- Wason, P.C. (1968), Reasoning about a rule. Quart. J. exp. Psychol., 20, 273-81.
- Wason, P.C. (1969), Regression in reasoning? Brit. J. Psychol., 60, 471-80.
- Wason, P.C. and Johnson-Laird, P.N. (1969), Proving a disjunctive rule. Quart. J. exp. Psychol., 21, 14-20.
- Wason, P.C. and Jones, S. (1963), Negatives: denotation and connotation. Brit. J. Psychol., 54, 299-307.
- Watson, G. and Glaser, E.M. (1964), Manual for the Watson-Glaser Critical Thinking Appraisal. New York: Harcourt, Brace and World.

- Watts, A.F. (1944), The Language and Mental Development of the Child.
London: Harrap.
- Wilkins, M.C. (1928), The effect of changed material on ability to do
formal syllogistic reasoning. Arch. Psychol., 16, 1-83.
- Woodworth, R.S. and Schlosberg, H. (1959), Experimental Psychology.
Fourth Edition. London: Methuen.
- Woodworth, R.S. and Sells, S.B. (1935), An atmosphere effect in formal
syllogistic reasoning. J. exp. Psychol., 18, 451-60.
- Woodworth, R.S. and Sheehan, R. (1965), Contemporary Schools of Psycho-
logy. London: Methuen.

and the validity of the corresponding argument as a whole, and an adequate grasp of, or adherence to, the 'logical task'.

In Chapter 2 the selection, on the basis of performance Forms of Section B of Valentine's Reasoning Tests, of a (PR) group of 'poor reasoners' and a (GR) group of 'good reasoners', matched on a composite measure of their (superior) academic attainment is described. Preliminary evidence relating to the difference between the groups is considered and the materials and procedure in the 'five types of statement' (5TS), experiment described. This experiment applies a form of Wason's 'card-turning task' to the three types of statement occurring most frequently in the criterion Valentine test, the universal affirmative in categorical form (A), the universal negative (E) and statements of the form, 'Only X's are Y's', and to two others incorporating logically important connectives, the universal affirmative in hypothetical form (Ah) and a universal-disjunctive (Ad).

The assumption underlying the 5TS experiment is that any difference between the groups in their grasp of any of the types of statement will be reflected in a difference in the success with which they perform the task with respect to that statement-type. Significant differences were found on A and E and on one aspect of the response to Ah, but not on E or on Ad. A progressive analysis of responses to different statement-types suggests the operation of other, unsuspected, factors and casts doubt on the validity of the use of the card-turning task as a measure of the relative difficulty of different statement-types.

In the later (4TS) experiment, incorporating modifications in the materials and procedure, it was possible to establish (1) the persistence of the differences on A and E over a period of 12 to 18 months and (2) the superiority of the GR group in learning to make the correct

Chapter 5 describes the use of Wason's 'construction task' to investigate the possibility that PR subjects have significantly more difficulty than their GR counterparts with the negative particle, an important element in all reasoning. The outcome is somewhat equivocal, there being a significant difference between the groups on the false-affirmative condition (which incorporates a single negative component, the 'semantic' notion of falsity) and when this condition is combined with the true-negative, but not on the false-negative which includes both true and negative components. This last result, it is suggested, may be due to the conspicuous instability of the response-time measure of difficulty in this case.

Chapter 5 also considers evidence (from scores on the E.P.I. about the extraversion and emotionality of the two groups, it has been argued that differences in either of these dimensions of personality may be partly responsible for difficulties with the negative particle. Although the differences are in the expected direction none is large enough to be significant on a two-tailed test. A small but significant difference in reaction time between scores on the N scale of the E.P.I. and latency of response to the true-negative condition when both groups of subjects are tested together lends, again in the absence of corroboration from the false-negative condition, rather uncertain support to the Eifermann hypothesis of a relationship between emotionality and slowness of response to negative statements.

A concluding chapter reviews the outcomes of the above experiments, considers an alternative interpretation of the 5TS results and suggests some possible points of departure for future research in this area.