PSYCHIATRY AND SCIENTIFIC METHOD:
PROBLEMS OF VALIDATING CAUSAL HYPOTHESES IN PSYCHOTHERAPEUTIC CONTEXTS

By

Michael Edward Dash

A thesis submitted to

THE UNIVERSITY OF LONDON

for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy, Logic

& Scientific Method

London School of Economics

January 2001

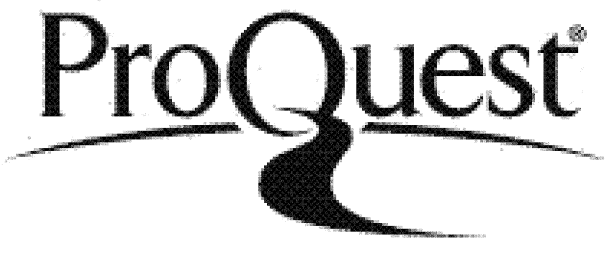
UMI Number: U615449

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.

UMI U615449
Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



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

THESES

F

7941

BRITISH LIBRARY OF POLITICAL AND ECONOMIC SCIENCE

844692

ABSTRACT

Adolf Grünbaum and others have criticised Freudian Psychoanalysis (FPA) methodologically, because of the potential inferential liabilities of testing causal psychoanalytic claims in the interview sessions. Also, Grünbaum has prescribed scientific experimentation (especially prospective group-comparison studies) as the best means of overcoming these methodological deficiencies. I argue, firstly, that the opportunities for making reliable inferences (including some causal ones) in psychotherapeutic interviews are better than Grünbaum canvasses, though I do this for a theoretically *weaker* form of psychodynamic psychology than FPA (i.e. General Psychotherapeutic Counselling, or GPC). Secondly, I argue that there are substantial problems with and limitations of experimental methodology, both in itself and when applied to test the specific kinds of hypotheses psychotherapists are interested in. Grünbaum and others fail to draw adequate attention to the generic problems of experimentation.

I argue that insofar as the acquisition of psychological knowledge is concerned, it is legitimate to distinguish two broad categories of inductive processes capable of providing it: (i) folk-psychological (or FP-) reasoning; and (ii) experimentation (including epidemiology). (By 'FP-reasoning' I mean the largely inherent capacity that human beings have for making inferences about the psychology of others or themselves, in folk-psychological terms.) I argue that FP-reasoning is in *some* respects inductively superior to experimentation. If this is correct, there ought to be no automatic methodological priority granted to experimentation in psychology and psychiatry.

Various topics related to the above are developed. For example: (i) some problems of the *practical application* to the psychotherapeutic domain of a principle of causal relevance (provided by Grünbaum) are examined; (ii) a sketch for a model of testing for causal relevance in the special case of insults (provided by Grünbaum) is criticised, and an alternative model is proposed; (iii) some general problems of justifying FP knowledge-claims are discussed, and sceptical attitudes towards the acquisition of folk-psychological knowledge are criticised.

CONTENTS

CHAPTER 1 *Introduction: Case-studies and Experiments* 12

1.1 Introduction and Overview 13

1.2 Case-studies and Experiments 21

1.3 FP-Reasoning and Experimentation 36

1.4 Background Themes to the Thesis 43

CHAPTER 2 *The Importance of FP-Reasoning and Some of Its Limitations* 47

2.1 FP-Reasoning is in Some Respects Inductively Superior To Experimentation 48

2.2 A Comparison Between FP-Reasoning and Experimentation 53

2.3 The Importance of FP-Reasoning for Mentalistic Understanding 60

2.4 Scientific Validation is Neither Necessary Nor Sufficient for the Recognition of the Truth of Some Non-Trivial Psychological Claims 62

2.5 An Historical Excursus – Freud's Dismissal of Experimentation 67

CHAPTER 3 *Opportunities for Case-study Causal Inference - An Example* 70

3.1 Reasons for Focusing on General Psychotherapeutic Counselling (GPC) 71

3.2 An Outline of General Psychotherapeutic Counselling 73

3.3 A Specimen GPC Case-study 78

3.31 Introduction 78

3.32 Preliminary Outline of the 'Tom' Case 81

3.33 The Reliability of the Data 83

3.34 Inferring Causal Conclusions – Some Examples 87

3.4 Concluding Remarks 99

CHAPTER 4 *Some Problems With Experimentation, and the Failure Adequately to Acknowledge Them* 102

4.1 Introduction 103

4.2 Research Designs 104

4.3 Problems Facing Group-Comparison Experimental Tests 110

4.4 Grünbaum and Eysenck Fail to Draw Adequate Attention to the Problems of Experimentation 125

4.5 Some Problems of Testing – An Example 140

CHAPTER 5 *Difficulties in Effectively* Applying *a Principle of Causal Relevance* 145

5.1 Introduction 146

5.2 Grünbaum's Standard Model of Testing for Causal Relevance (SMCR) 148

5.3 The Examples 150

5.4 Can the Informal Application of the SMCR Test α_1 Effectively? 153

5.5 Can the Synchronic Experimental Application of the SMCR Test α_1 Effectively? 158

5.6 Can the Diachronic Experimental Application of the SMCR Test α_1 Effectively? 168

5.7 Discussion 171

CHAPTER 6 *The Scientific Legitimacy of Seeking Individualised Knowledge – A Response to Cartwright* 175

6.1 Introduction 176

6.2 A Counterexample to Cartwright's Objection 178

6.3 Clarifications Relating to C1 and C2 182

6.4 The Need to Be Open About the Failure of Scientific Testing 189

CHAPTER 7 *Problems of Justifying FP Knowledge-Claims* 194

7.1 Introduction 195

7.2 Scepticism and Antiscepticism With Regard to FP Knowledge 199

7.21 Scepticism 199

7.22 Antiscepticism 208

7.3 Grünbaum's Account of FP Causal Relevance Evaluation for the Case of Insults 214

7.4 The Single-Case Model for Insults 219

7.5 Postscript 225

CHAPTER 8 *Conclusion* 230

NOTES 235

REFERENCES 252

ACKNOWLEDGEMENTS

This thesis is the outcome of study and research conducted in the Department of Philosophy, Logic and Scientific Method at the London School of Economics.

I am deeply grateful to those who were my teachers when I was a Master's student – it was they who provided an apprenticeship in logic and the philosophy of science for the Ph.D. that followed. Several staff members were primarily responsible for my training in these areas at the time, and I am especially grateful to them: Colin Howson (Deductive Logic, Probability Theory), Peter Urbach (Induction and Scientific Method), the late John Watkins (Philosophy of Science, Epistemology) and John Worrall (Deductive Logic, Philosophy of Science, History of Science). I also learned much that was of value from Helena Cronin (Neo-Darwinian Evolutionary Theory), David-Hillel Ruben (Philosophy of Action) and Elie Zahar (History of Science). Nancy Cartwright joined the Department soon after I began my Ph.D., and my huge debt to her is recognised below.

Apart from my gratitude to these members of staff for all they have taught me, I can sincerely say that all of them will have an enduring place in my affections.

Moving on to the present Ph.D. thesis, several persons have been especially important for my writing it, and for the form and content it has taken. I should like to thank each of them individually:

John Worrall

First and foremost, I should like to thank John Worrall who offered to be my Ph.D. supervisor. With almost saintly patience Professor Worrall has watched my original project (which was to have been on the philosophy of organic medicine, not psychiatry) evolve through many stages. The freedom granted me both by him and the Philosophy Department to allow my interests to grow and change has been greatly appreciated.

Throughout my time at the L.S.E. Professor Worrall has served as a model for my efforts to master the Philosophy of Science and Logic. However poor my accomplishments in these fields, they would have been a lot worse had it not been

for his example and teaching. He made deductive logic – a subject I found difficult – comprehensible and, dare I say it, simple. He also changed my outlook on the Mind-Body Problem (which, for what it was worth, had been a form of methodological dualism) to a kind of Monistic physicalism. Overall, he has assisted my learning and my efforts to arrive at a viewpoint of my own in more ways than I could recount

During the writing of my thesis Professor Worrall has been a relentless critic. But his criticisms, however difficult to respond to at the time, always led me to think more carefully about what I had written, to identify weaknesses that had not been apparent, and to clarify ideas that had not been expressed well enough. I doubt that Professor Worrall will wholly approve of the result. However, he has played a central role in helping me to bring my thesis to its present state.

Nancy Cartwright

Secondly, I should like to thank Nancy Cartwright, who was my second Ph.D. supervisor.

My first encounter with Professor Cartwright's views – which came from reading her book *How the Laws of Physics Lie* in 1992 – were, for me, nothing short of an intellectually liberating experience. There was much in her views on science that I felt to be important, but that had not been central (if recognised at all) in the work of philosophers of science I had studied up to that time. For example: her emphasis on practice, and on the empiricism that goes hand in hand with practice; the centrality of causal reasoning; and her criticism of the view that the abstract mathematical laws of physics are fundamental. In trying to find a perspective on science for which I feel some degree of intellectual conviction, I have been strongly influenced by Cartwright's views. I am also very grateful to Professor Cartwright for some specific comments she has made to parts of my thesis (see especially chapter 6 and section 7.5). These have stimulated my thought and have led to at least one problem for which I currently have no adequate response (7.5). Overall, it has been a privilege to have had Professor Cartwright as one of my supervisors.

Adolf Grünbaum

Although I have had the pleasure of meeting Professor Grünbaum when he has visited the L.S.E., I have never been formally taught or instructed by him. In spite of this, I have felt it essential to include him here: if for no other reason than

that it would be *inaccurate* to omit him from the list of persons who have been most responsible for the content of this thesis.

Firstly, it was Professor Grünbaum's [1984] book *The Foundations of Psychoanalysis* that made possible the present thesis. I began reading Grünbaum's book in earnest in 1992, and it has been the primary inspiration and stimulus for my thesis.

Secondly, it is no exaggeration to say that, through reading Professor Grünbaum's writings on psychoanalysis I have learned as much about the philosophy of science as I have through any of my formal teachers (and this is as a bonus to learning about Freud). Whilst working in this field I have certainly thought of Grünbaum as one of my teachers – as one of the persons from whom I have had the most to learn.

I am in agreement with the prevailing view that Grünbaum has been the most important philosophical commentator on Freud. But, more than this, I think that through his work in this area he has made possible the beginnings of the development of the philosophy of psychiatry as a sub-branch of the philosophy of science. How this field will grow in the future cannot be foretold, but Grünbaum has set the standard of analysis and of rigour that those who follow will need to emulate. I should like to emphasise the latter point given that, in my thesis, I criticise Grünbaum's views in various places. Integrity demands that we follow our intellectual intuitions as best we can, even when these lead us against the views of those we admire (and, of course, my arguments may be full of flaws). In spite of my criticisms or disagreements, the truth is that I have a great admiration and respect for Grünbaum's work.

James Hopkins

I am very grateful to James Hopkins for discussing various aspects of the philosophy of psychoanalysis with me during the early stages of my thesis. Dr. Hopkins made me aware of the importance of the philosophy of mind for philosophical work in psychiatry and psychotherapy. It is primarily through him that I developed an interest in (so-called) folk-psychology, a topic which came to play an important role in my thesis. The specific ways in which I have developed arguments relating to folk-psychology differ, I think, markedly from his own in various respects. Nevertheless, had it not been for Hopkins's influence I may never have

come to appreciate the relevance of folk-psychology for understanding in psychotherapy at all.

K. W. M. Fulford

Professor K. W. M. Fulford has probably done more than anyone else in Britain to open up a fertile exchange between philosophers and psychiatric professionals. It was my good fortune to have had the opportunity to attend meetings of the Philosophy Special Interest Group of the Royal College of Psychiatrists (London) in the early 1990s, and to attend a course for psychiatrists and philosophers given jointly by King's College (London) and the Royal College. Professor Fulford was responsible for setting up the Special Interest Group, for organising the course, and for going on to found the journal 'Philosophy, Psychiatry, and Psychology' which has become an organ for interdisciplinary work in the area. The interests I developed whilst attending these meetings and participating in the discussions were central to my decision to conduct research in this area. Consequently, I have a lot to thank Professor Fulford for.

'Tom'

Last, but not least, I should like sincerely to thank the person referred to as Tom, some aspects of whose personal life I discuss as a case-study in chapter 3. Without this material, drawn from real-life, I would not have been able to think realistically about the psychological and methodological issues involved. I am grateful for the candour with which Tom has communicated issues of personal emotional significance. It is a testimony to the degree of trust between us that this has been possible.

Apart from the above persons whose influence on my thesis has been the greatest, I should also like to thank those research students who were my colleagues in the early-mid 1990s who stimulated my thought through discussion: Samet Bagçe, Tim Childers, Robin Hendry, Andrew Powell, Stathis Psillos, Towfic Shomar and, especially, Kolapo Abimbola. Thanks also to Gabriel Segal for his critical point (with Nancy Cartwright) on my 'Single-Case Model' (see section 7.5).

Finally, I should like to thank Liz Allmark for her quality typing of the first three chapters.

For three wonderful and very special children:

Rosie, Damien and Lucy

With my love

CHAPTER 1

INTRODUCTION:

CASE-STUDIES AND EXPERIMENTS

1.1 INTRODUCTION AND OVERVIEW

The *specific* aim of this thesis is to examine some of the problems that arise in connection with the testing and validation of hypotheses in psychotherapy. The focus will be predominantly on causal hypotheses. The thesis also has a more *general* aim, which will be presented below.

There can be no doubt about the considerable extent to which psychotherapists seek causal knowledge. In a standard series of psychotherapeutic sessions the therapist will make a large number of tacit or explicit causal hypotheses and will arrive at a large number of causal conclusions, even if these are held only conjecturally. For example, inferences may be made about the likely causal relevance of various past events, concurrent circumstances, or mental states for (other) mental states, behavioural attributes or even the personality of the client. Causal hypotheses are not, however, the only ones of interest to psychotherapists. Also of interest are what might be called 'descriptive' hypotheses. These include hypotheses which are descriptive of cognitive, emotional or behavioural states (e.g. John believes that p; Sue fears X; Tom is socially inactive) when questions about the causes of the states are either not the primary interest at the current time, or are not asked at all. Descriptive hypotheses also include hypotheses about possible events or experiences in the client's life at some stage in the past – e.g. that Jane had a traumatic experience of type Z when she was a child. Whereas descriptive hypotheses can, in principle, often be answered by the accurate provision of the relevant information by the client (or by making an observational judgement of him/her in behavioural cases), causal hypotheses make claims about a relationship of causal relevance between two factors or variables (the so-called independent and dependent variables). It is the difficulties of evaluating whether or not such a relationship holds in a given case that lends causal hypotheses their peculiar problematic status.

The present thesis can most easily be introduced by way of a contrast with (and as a partial response to) the work of several authors, each of whom has provided a methodological and epistemological critique of psychoanalysis. The authors concerned are Adolf Grünbaum, Hans Eysenck and Edward Erwin (I shall focus predominantly on Grünbaum's work). Before stating the two broad themes which relate this thesis to aspects of their work, some preliminary points of clarification need to be made to avoid potential misunderstanding.

Firstly, the above authors are primarily interested in evaluating Freudian psychoanalysis (FPA) – i.e. its theory, particular claims, practical methodology and epistemological foundations. I shall not be doing this. Even though I use a few Freudian examples, I have no special interest in FPA or its evaluation. Insofar as I touch upon Freudian themes it is because they overlap with more general epistemological or methodological issues which are my concern. Most of my examples will come from what I shall call General Psychotherapeutic Counselling (GPC), or even from the 'commonsense' psychology of everyday life termed 'folk-psychology' (FP) in the philosophical literature (e.g. see Dennett [1987] 7).

Secondly, this thesis is not intended to be a systematic or point for point response to the above authors' views or comments on psychoanalysis. Not only are there significant differences in their general viewpoints (e.g. Eysenck is more extreme in his advocacy of experimentation than Grünbaum), but the sheer number of topics which they deal with would have made any systematic response to their views excessively lengthy and disputatious. Even though I do present discussions or arguments in response to some specific points made by them, this is selective and is used mainly to further my own agenda of developing and presenting my own position in the field. The reason why I have taken these authors as a 'group' is because, in spite of differences between them, they all express a similar outlook (see below). Moreover, even though I have (with all due respect) treated the work of these authors (especially Grünbaum) as a foil against which to develop my own ideas, I do not by any means disagree with them on all (or even some basic) issues.

My disagreement with the above authors is primarily at a very general level – it concerns their emphases and also, importantly, the things they leave out. It

is the general aim of this thesis to assist in redressing this, and to restore what I believe is a more balanced overall perspective to psychotherapeutic epistemology (this is the general aim alluded to in the opening paragraph). In particular:

> (i) The above authors do not make a strong enough positive case for the acquisition of knowledge (including causal knowledge) in psychotherapeutic case-studies. (This is even if, for present purposes, we ignore the strong and distinctive claims of FPA and confine ourselves to GPC.)
>
> (ii) They are insufficiently critical of the inferential powers and effective applicability of experimental methods, especially in regard of the effectiveness of such methods for testing many of the hypotheses in which psychotherapists are interested.

My attempt to counter each of these tendencies makes up the bulk of the central part of this thesis (chapters 2-5).

In chapter 2 I examine some of the features of the inductive modality which is utilised when we make informal inferences about human psychology in everyday life (I call this modality 'FP-reasoning'). While I acknowledge that there are considerable limitations to FP-reasoning, I also attempt to illustrate some of its inferential powers and epistemic virtues (this is something Grünbaum, Eysenck and Erwin do not do at all).

In chapter 3 I present a favourable case (albeit, as I see it, a modest one) for the informal testing and validation of some types of causal hypotheses in uncontrolled GPC case-studies. This chapter includes a case-study of my own (the 'Tom' case).

Chapters 4 and 5 are devoted to a critique of experimental (or controlled) methodology as applied to the psychotherapeutic domain. In chapter 4 I discuss some of the problems facing experimental methods and go on to argue that Grünbaum and Eysenck do not draw adequate attention to these. In chapter 5 I

critically discuss some problems of applying in practice a model of testing for causal relevance advanced by Grünbaum.

In chapter 6 I respond to a criticism pointed out by Nancy Cartwright which relates to my discussion in chapter 5. In essence, I argue (against the criticism) that it is the legitimate concern of science to be interested in testing/validating singular (or 'individualised') causal claims and that there is no principled restriction to general (or 'generic') ones. One of the morals to be drawn from this is that scientific methods of testing cannot be excused from being judged as failures simply on the grounds that the causal hypothesis at issue is singular (as opposed to generic).

Chapter 7 is largely autonomous. It examines various problems connected with justifying folk-psychological (i.e. FP) knowledge. This topic is seen as central and, in a sense, as foundational given that so much alleged psychotherapeutic knowledge begins with everyday 'commonsense' psychological understanding or builds on it. I begin by contrasting sceptical and anti-sceptical attitudes towards FP-knowledge. I then critically examine (for a restricted case) a model of justifying FP-knowledge put forward by Grünbaum, and I propose a different model of my own. The chapter concludes with my recognition of a very interesting (and unresolved) problem facing my model pointed out by Nancy Cartwright and Gabriel Segal.

One piece of terminology that it will be useful to introduce at this stage is the following classification scheme for different types of hypotheses in which psychotherapists will be interested[1]:

Category I hypotheses: hypotheses of everyday 'commonsense–' or folk-psychology (FP). For example: 'John believes that p', 'Mary is jealous of Sue', 'Bill wishes to go to Hollywood', 'the catching alight of the frying pan caused Jill to become alarmed', 'people tend to put on warm clothing when they feel cold'.

Category II hypotheses: FP-like psychotherapeutic hypotheses. These are hypotheses which, in terms of theoretical content, are very similar to those of FP, but which tend to be considered primarily in psychotherapeutic sessions of a theoretically weak 'counselling' kind (such as GPC). Grünbaum in effect refers to a category II hypothesis when he quotes Rosemarie Sand's example of a female patient whose analyst hypothesised that "[her] low self-esteem – as manifested by her expectation of contempt from [the analyst] – was caused by her lifelong awareness of her father's vilifications of her" (Grünbaum [1993] 112). The hypotheses T1-T4A which I deal with in connection with the 'Tom' case (chapter 3) are also examples of category II hypotheses.

Category III hypotheses: these are theoretically strong (or otherwise distinctive) hypotheses from one or other of the doctrinal psychodynamic schools. They include the hypotheses which Grünbaum takes to be the core ones of Freudian theory (Grünbaum [1993] 110-111 – see footnote 5). Also included would be, for example, Freud's claim that a "contrary" psychodynamic relation holds between neuroses and sexual perversion (neuroses require repression; perversion is marked by regression but the absence of repression – Freud [1917](c) 388, 393; Freud [1917](d) 405).

Category IV hypotheses: these are hypotheses which have been formulated and tested as part of scientific psychology and medicine (or one of the established natural sciences), but which may be of background relevance to psychotherapists and may enter into their clinical judgements. The methods of testing hypotheses in category IV will almost always be 'extraclinical' (i.e. experimental or epidemiological) and in keeping with orthodox scientific procedures. Examples include genetic or biochemical hypotheses about schizophrenia and manic-depressive illness; Behaviouristic (i.e. conditioning) theories of neurosis; statistically based theories of personality-type etc.

A second terminological point concerns my use of the word 'experiment' (or 'experimentation'). Except where a specific distinction needs to be made between them I shall, for the sake of brevity, use the word 'experiment' to cover both experiments and epidemiological studies[2] (although I shall sometimes use both terms together for emphasis).

There is in fact no absolutely clear-cut distinction between experiments and epidemiological studies although, in general, experiments tend to involve a greater degree of control over extraneous variables and also often have greater scope for the manipulation of the independent variable(s). Randomized Clinical Trials (RCT's) are, for example, regarded as amongst the methods available to epidemiologists but are also referred to as experimental (Backett and Robinson [1992] 204). What is characteristic of both kinds of study (as compared to informal or 'commonsense' inference) is that they involve specially introduced strategies which serve specific methodologically normative ends. For example, a particular method of collecting data may be used to increase its objectivity or its 'representativeness' of the population; external controls may be applied to help eliminate the influence of extraneous variables; and special analytical or statistical techniques may be employed to assist in making an inference.

A third terminological point is that, for the sake of conciseness, I shall use the letters 'FP' to stand both for 'Folk Psychology' and 'Folk Psychological'. Whether the substantival or adjectival sense is intended should be clear from the context.

A final but important preliminary point concerns the restriction of my interest when dealing with psychotherapy in this thesis to its epistemological aspects. Psychotherapy has two sides to it[3]. Firstly, there is an epistemological side which is concerned with the attempt to acquire knowledge about an individual's psychology

or psychopathology (or to obtain evidence for more general claims). Secondly, there is a therapeutic side, which is concerned with bringing about a therapeutic effect in the client (for example, relieving symptoms of anxiety or depression, changing incapacitating false beliefs, changing maladaptive behaviour, remedying difficulties in interpersonal relations, increasing self-esteem etc.). These are clearly distinguishable aspects[4]. Conceivably, one could bring about a therapeutic effect in a client in a psychotherapy session without any adequate understanding of how it was achieved. Equally, it is conceivable that one could genuinely learn something about a client's psychology in a psychotherapy session without bringing about any therapeutic benefit (or perhaps even with the consequence of making the client's condition worse). In this thesis I shall restrict my concerns entirely to the epistemological aspects of psychotherapy. That is, I shall be concerned with psychotherapy only to the extent to which the psychotherapeutic situation (of, typically, a therapist and his client) provides an opportunity for acquiring knowledge or testing hypotheses. I shall not be concerned with questions about the possible therapeutic effect of psychotherapy or with the evaluation of therapeutic claims. To this extent the very terms 'psycho*therapy*' and 'psycho*therapeutic* session' are misnomers for my concerns. The terms can, however, be retained because conventional usage allows inclusion of hypothesis-testing or knowledge-acquiring capacities.

If justification is needed for the above restriction the following might serve:

The problems of psychotherapeutic epistemology are considerable in their own right and not everything that deserves attention can be tackled at once. Focusing on a restricted set of problems does not imply that others which are neglected are less worthy. The ultimate goals of psychotherapy and psychiatry are aesthetic, normative and practical – i.e. improvement of the mental health, well-being and social adjustment of individuals (and, in my view, at least a contribution towards helping to create a better society through these). Although the attainment of these goals might be possible to some degree without understanding, any effective and systematic attempt to achieve them (if problems arise) must be based on

knowledge. Many different types of study, involving different theoretical approaches and methods, might conceivably contribute towards providing the requisite knowledge (i.e. of human psychology, psychopathology and of how to bring about therapeutic change). The epistemological resources of psychotherapy might be able to make their own distinctive contribution towards facilitating the acquisition of that knowledge. This, if nothing else, warrants their being carefully examined and evaluated.

1.2 CASE-STUDIES AND EXPERIMENTS

Grünbaum, Eysenck and Erwin all draw attention to the inferential and epistemic pitfalls of psychoanalytic case-study inference. Grünbaum says, for example:

> "It is one central thesis of this essay [i.e. Grünbaum's 1984 *magnum opus*] that the clinical psychoanalytic method and the causal (etiologic) inferences based upon it are fundamentally flawed epistemically . . ."
>
> (Grünbaum [1984] 124)

Eysenck and Erwin express similar views (e.g. Eysenck [1953] 228-229; Erwin [1986] 227). Grünbaum is especially thorough in his analysis, examining factors which tend to 'contaminate' the clinical data. These include suggestion by the analyst (see e.g. Grünbaum [1984] 130-131, 212-215, 244-245, 264) and the fallibility of the patient's memory (op. cit. 242-244). However, Grünbaum argues that even if the clinical data were uncontaminated, uncontrolled case-studies, as used in psychoanalysis, would lack the inherent ability to test effectively the "major" or "cardinal" causal Freudian hypotheses[5] (op. cit. 128; see also Grünbaum [1993] 3, 110).

All of the above authors also view experimental or epidemiological methods as the principal means of improving upon or rectifying the limitations and pitfalls of psychoanalytic case-study inference. The advocacy of such methods constitutes, in fact, their main methodological prescription. The use of experimental (or controlled) studies is seen by them as the best means of bringing the testing of psychoanalytic hypotheses into the fold of *bona fide* scientific practice. As evidence for this consider the following quotations:

Grünbaum

> "In view of my account of the epistemic defects inherent in the psychoanalytic method, it would seem that the validation of Freud's

cardinal hypotheses has to come, if at all, mainly from well-designed *extra*clinical studies, either epidemiologic, or even experimental . . ."
(Grünbaum [1984] 278; emphasis in original. For a slightly stronger statement of the same view see Grünbaum [1986] 228)

"Being replete with a host of etiological and other causal hypotheses, Freud's theory is challenged by neo-Baconian inductivism to furnish a collation of positive instances from *both* experimental and control groups, if there are to be inductively *supportive* instances.

. . . to this day, analysts have not furnished the kinds of instances from controlled inquiries that are *inductively required* to lend genuine support to Freud's specific etiologies of the neuroses."
(Grünbaum op. cit. 280; emphases in original)

Eysenck

"It is for [reasons of avoiding the bias and error which are due to psychoanalytic training and practice] that we have laid stress on experimental studies; these alone can have the objectivity, the detailed description of conditions and results, and the requisite statistical elaboration to make proper judgement possible. The majority of psychoanalytic writings are lacking in these essential safeguards, and hence are not of much value for those searching after truth."
(Eysenck and Wilson [1973] 8)

Erwin

"Grünbaum's work . . . creates a strong presumption that existing clinical data provide little, if any, support for Freudian theory. If this presumption is to be overcome, an analyst must show how the epistemological difficulties are to be resolved. If that cannot be done, then the best bet is to look for support in (non-clinical) experimental studies."
(Erwin [1986] 188; references omitted)

" . . . I disagree with the sceptic who denies the very possibility of uncontrolled clinical confirmation of causal hypotheses[6]. . . . However, I remain sceptical about both the actuality and possibility of non-experimental confirmation of psychoanalytic causal views. I agree, in short, that the vindication of Freudianism requires sound and extensive experimentation."
(Erwin [1988] 215)

A sympathetic reading of the views of (at least) Grünbaum and Erwin requires us to conclude that they are not claiming that there definitely are (or, in the future, will be) experimental methods which can effectively test the causal psychoanalytic claims in question in every case, and that these methods simply need to be applied. Rather, we should see them as upholding a conditional claim to the effect that if any method is going to be able to test effectively the core causal claims of psychoanalysis it will be an experimental or epidemiological method (and not the uncontrolled psychoanalytic case-study method). Nevertheless, this qualification should not alter our appreciation of the confidence with which they canvass experimental and epidemiological methodology as a potential remedy for the liabilities of the psychoanalytic case-study.

My attitude toward the above stance comprises both: (A) a (modestly) more favourable attitude, overall, towards uncontrolled case-studies; and (B) a much more critical attitude towards experimentation. Let us consider these in turn:

(A) Whereas the above authors are entirely justified in pointing out the liabilities of uncontrolled case-study inference, especially for FPA, I feel that a somewhat stronger case can be made for such inference. Since my defence will be given for GPC case-studies (and not for distinctively FPA ones) some justification for this needs to be given. After all, it could be objected that:

(i) Even if a successful defence of the weaker (GPC) case can be given, this does nothing to offset the legitimacy of the criticisms which Grünbaum brought to bear against the stronger (FPA) case, and that these criticisms remain unresolved.

(ii) Grünbaum would have little difficulty in accepting that there are stronger grounds for accepting conclusions drawn in GPC case-studies than in FPA ones. Indeed, Grünbaum explicitly acknowledges such a difference

(Grünbaum [1993] 112; also of relevance is Grünbaum [1984] 229-230). Consequently, it could be objected that I defend - or make a positive case for – something which Grünbaum might in any case have no difficulty in agreeing with (whilst ignoring the difficulties inherent in the stronger, FPA, case).

Regarding point '(i)': I am, for present purposes, content to let matters rest as the objection states. My aim is to show up uncontrolled case-studies in a (modestly) more favourable light than Grünbaum, Eysenck and Erwin have done[7]. This is not inconsistent with accepting that the classical Freudian variant of such case-studies carries substantially greater liabilities than the GPC version. Equally, it is not inconsistent with a (provisional) acceptance of Grünbaum's claim that FPA case-studies are ineffective for providing crucial tests of (what he takes to be) the core causal hypotheses of Freudian theory (see footnote 5).

Regarding point '(ii)': my criticism is not that Grünbaum does not recognise that we might sometimes be able to draw reliable conclusions from uncontrolled case-studies (e.g. see Grünbaum [1984] 245, 259-260, 265-266); or that he does not acknowledge a greater capacity for this in GPC studies than in FPA ones (op. cit.). Rather, it is that he does not build on his recognition of the mere possibility of these things and supply an actual positive account of them. The main import of my criticism is not, however, to blame Grünbaum for what he has not done. It is, rather, directed impersonally against the absence of an account which would show the inferential capabilities of case-studies in their best light.

To add to the above, it should be noted that there is no strict demarcation between GPC and FPA case-studies. This applies to the kinds of hypotheses considered in the two types of case-study, the kinds of data that could crop up in them, and the general strategies of inference used. It would, I think, be more accurate to say that there is a spectrum of (uncontrolled) psychotherapeutic - or psychodynamic – case-study types. Towards one extreme are case-studies which are strongly interventionist, make considerable use of special techniques (such as free-association, or an epistemically facilitating use of transference phenomena), and

involve theoretically strong hypotheses (this is the type of case Grünbaum focuses on in dealing with classical FPA). Towards the other extreme are minimally interventionist case-studies relying on voluntary disclosures by the client in which the hypotheses are essentially FP-like (this is the type I shall focus on). However, I do not think that there is any insurmountable barrier to non-FP-like data arising in a GPC case-study. Suppose for example (just for argument's sake), that an adult male client in a GPC session had veridical memories (of when he was an infant) of specific anxieties of the kind that Freud refers to as castration anxiety, and that he volunteers this information to the psychotherapist. This might lend strong support to Freud's (descriptive) claim that castration anxiety is a psychological reality (for, at least, some male infants), even though the supporting evidence arose in a GPC case-study. That any infants experience a specific anxiety of this kind is certainly not part of FP (even though it should be possible to describe the main phenomenological features of such anxiety – if it exists – in terms of FP concepts). If data of the above kind were volunteered and there were good grounds for believing the testimony to be credible, it would have to be accepted, at least provisionally. That such a circumstance could actually occur in a GPC session cannot, moreover, simply be ruled out by fiat. If it were objected that it is not possible because Freudian theory proclaims that such experiences, if genuine, will be repressed (and require psychoanalytic technique to extricate them), then I think the appropriate response would be to say that this objection makes the error of putting the cart of theory before the horse of evidence (or observation). On this latter point, I believe that psychotherapists (insofar as they perform an epistemological function) should be, first and foremost, good empiricists and observers; and should make themselves receptive to whatever data presents itself, however unexpected.

Furthermore, whereas Grünbaum is perfectly entitled to restrict himself to classical FPA case-studies and to (what he takes to be) the core causal claims of Freudian theory, it should be borne in mind that this is not entirely representative of the epistemic situation in psychotherapy (or psychodynamic psychology). Not only – as we have seen – are there 'milder' forms of case-study, but there are also many hypotheses of interest – both causal and descriptive – which

fall outside the range of the core causal ones of Freudian theory defined by Grünbaum. Consequently, I think it would be unjustified - or at least premature - to reach an encompassing negative verdict on the epistemic capabilities of case-studies simply on the strength of (let us suppose) Grünbaum's success at demonstrating their inadequacy for the examples he focuses on. To do the latter would risk throwing out the baby with the bathwater.

(B) Regarding the above authors' advocacy of experimentation, I believe that a much more critical attitude towards it than they offer is deserved. This applies to each of the following[8]:

> (B)(i) The question of whether experimental methods can, indeed, provide critical and effective tests for the kinds of hypotheses that psychotherapists are interested in;
>
> (B)(ii) Various theoretical and practical problems of a general kind to which experiments are subject;
>
> (B)(iii) More fundamental issues.

Grünbaum, Eysenck and Erwin do analyse and critically discuss various experimental studies bearing on psychoanalysis and related topics (e.g. Grünbaum [1984] 202-205, [1986] 269-270; Eysenck and Wilson [1973]; Erwin [1980] 446-454, [1986] 191-199, [1988] 215-221). This includes evaluating experimental results, drawing conclusions from them, and commenting on various aspects of experimental design. Significant though these discussions are they do not go far enough in addressing the difficulties encountered in (B)(i) – (B)(iii), above. Let us consider these in turn.

Regarding '(B)(i)', I believe that we need to consider not only whether experimental methods can effectively test Freudian (or other theoretically strong psychodynamic) hypotheses (i.e. in category III), but also whether they can effectively test (comparatively) weak hypotheses of GPC, or even FP (i.e. categories I and II). Indeed, although I discuss an example bearing on the first of these sub-problems (section 4.4), I have been more interested in the second and devote more attention to it.

The (former) problem of whether there are effective experimental tests of distinctively Freudian hypotheses (and not just of reconstructed hypotheses which in some superficial respects resemble the Freudian ones) is a longstanding one. Erwin recognises it when he says:

> "One of the main complaints [of psychoanalysts] is that in trying to devise strict controls, the experimenter is forced to test propositions that bear only a faint resemblance to Freudian hypotheses."
> (Erwin [1986] 186)

Grünbaum is also clearly aware of this problem (e.g. Grünbaum [1984] 101-102; [1980] 368, 373; and the passage from Kline [1981] 437 quoted in Grünbaum [1986] 270). However, whereas these authors recognise the problem, not enough is done to make a convincing (positive) case that *bona fide* Freudian claims can be effectively tested, experimentally (see e.g. Erwin op. cit. 190-191; Grünbaum [1986] 268). Provision of a stronger case for this would help to rebut the sceptic, and would make the conditional nature of these authors' advocacy of experimentation for this purpose (see page 23) seem more than a promissory hope. Critical reservations about the extent of the experimental testability of psychoanalytic claims are in order. Edelson, for example, has said:

> ". . . Grünbaum seems to me to be uncharacteristically ingenuous in recommending carrying out epidemiologic and experimental studies to test psychoanalytic hypotheses (while also maintaining that it is fruitless to turn to clinical data to test these hypotheses), without at the same time emphasizing:
> (1) the exceptional difficulties posed even in using such studies to test nonpsychonalytic hypotheses in the clinical realm;

> (2) how much greater the difficulties are in psychoanalysis, where causal variables are often intrapersonal (e.g., fantasies) rather than situational (e.g., parental behaviour) and where it is impossible, unfeasible, or unethical to mobilize the phenomena of interest (e.g., intense and, for the subject, highly objectionable sexual or aggressive impulses) in anything like the relevant intensities outside the clinical situation."
>
> (Edelson [1986] 232; and repeated in [1988] 275)[9]

Regarding the experimental testing of category I and II hypotheses, this is a topic which, to the best of my knowledge, has received little or no attention in the philosophy of science literature and which might even be viewed as undeserving of serious intellectual attention. I strongly disagree with the latter attitude. I accept that a practising experimental research scientist will not be concerned with it (and has little need to be). Experimental scientists tend to take the 'commonsense' background knowledge on which we rely (whether FP or folk-physics) for granted and/or as unproblematic – and, in any case, they have other tasks to attend to. However, for the epistemologist and philosopher whose duty is to examine the foundations of knowledge and of methodology there can be no excuse for overlooking the issue. Perhaps the most widespread view is that it is simply a matter of 'scientific commonsense' that experimental methods are not suited for testing FP hypotheses or FP-like hypotheses of GPC (especially if the hypotheses are singular). This viewpoint has to be rejected for the simple reason that it explains nothing. If it is the case that experimental methods cannot be applied effectively to test certain categories of FP or GPC hypotheses then we need to know why. Whereas I do not presume to be able to provide a fully satisfactory answer, I do consider the question (see sections 3.4, 5.5, 5.6, 5.7, 6.4, 7.21).

I appreciate that by shifting ground onto this aspect of experimental application (or lack of it) I am no longer addressing the question of whether experiments can be used to test Freudian (or other theoretically strong psychodynamic) hypotheses, and yet that it was in relation to the latter issue that I gave an account of Grünbaum, Eysenck and Erwin's prescriptive advocacy of experimentation. It might hardly seem pertinent (or fair) to decry their advocacy of experimentation for testing category III hypotheses by pointing to the inefficacy of experimentation for testing category I or II hypotheses (if that turned out to be the

case). Their methodological prescription, after all, was not that there should be a thorough experimental testing of category I or II hypotheses. However, it could be argued – hypothetically or *prima facie* – that since FP or FP-like GPC hypotheses are theoretically weaker and/or observationally more accessible than FPA ones, if experimental methods cannot be used effectively to test (many of) the former, it is misplaced even to expect them effectively to test the latter. On the basis of this it might be felt that an examination of whether experiments can test the weaker cases (and, if not, why) was a prerequisite of evaluating their adequacy for testing the stronger ones. I am not claiming that the above argument is a sound one[10] – only that it could, *prima facie,* be used to counter any charge of the irrelevancy of giving prominent attention to the experimental testing of category I and II hypotheses. Any fundamental examination of the scope of effective experimental application must (ultimately) engage not only with the question of what is the (positive) range of that application but also under what conditions (or for what hypotheses) effective experimental testing breaks down – and why. A major reason for giving attention to whether experiments can test hypotheses in categories I and II is to permit the beginnings of an exploration of that possibility, at least for psychology.

In any case, it is not true that the experimental testing of FP-like hypotheses has no bearing on any of, at least, Grünbaum's views or interests. In a response to Thomas Nagel's comments that experimental methods are largely inappropriate and/or inapplicable for testing FP hypotheses (Nagel [1994] 35) Grünbaum replied:

> ". . . Nagel regards as "telling" against me [the] question "How is common sense psychology tested?" In the case of its ubiquitous causal hypotheses, such as that insults anger or humiliate people, good tidings create joy, or that people tend to put on protective clothing because they feel cold, I reply: To warrant that a factor of sort X (such as being insulted) is causally relevant to a kind of outcome Y (such as being angered or feeling humiliated) in a reference class C, evidence is required that the incidence of Y's in the subclass of X's is *different* from its incidence in the subclass of non-X's. Nagel speciously pleads the idiocy or impossibility of statistical confirmation of such general hypotheses, irrelevantly lampooning statistical evidence for their *individual application* to particular instances in either common sense psychology or psychoanalytic theory."
>
> (Grünbaum [1994](a) 54; emphases in original)

Grünbaum therefore clearly does have a viewpoint about the ability of experimental or epidemiological methods to test at least some FP (i.e. category I) hypotheses effectively. In the case of the kinds of general, causal FP hypotheses that he mentions he clearly believes that they can be statistically confirmed (and, by implication, tested). (He criticises Nagel for what he sees as the latter's misattribution of statistical evidence to the individual case.) Consequently, when I take up the question of the experimental testability of category I or II hypotheses it cannot justifiably be objected that this has no bearing on any views held by Grünbaum (even though it does not directly address his primary concern, which is to use experimental or controlled studies to test FPA claims of category III).

Regarding '(B)(ii)' (i.e. problems of a general kind which experiments face), these are discussed in sections 4.3, 4.4 and 4.5, so there is no need to elaborate on them here.

Regarding '(B)(iii)' (i.e. more fundamental issues), there are many topics which could, conceivably, fall into this category. For example, philosophical justifications of causal inferences in experimentation, or more technical issues concerning the bases of probabilistic or statistical inference. What I have in mind will, however, be different from the latter. I have been especially concerned with attempting to compare experimentation with non-experimental (or informal) reasoning, particularly with a view to questioning the widespread assumption that experimentation is always or is in every respect (or is even 'obviously') epistemically superior. Consequently, I have been interested in such questions as: 'What are the (theoretical and practical) limits of the effective application of experiments for testing psychological hypotheses?'; 'Can using experiments actually restrict the type of knowledge we can acquire (in the psychological domain)?'; and especially 'Are there any respects in which non-experimental inference is actually

advantageous over experimentation in psychology (and, if so, what are these ways)?'.

One should not, of course, prejudge what the outcome of asking questions such as these will be. However, not to ask them seems to me to grant too much by way of automatic (and unexamined) right to experimentation. It also seems that Grünbaum, Eysenck and Erwin fail to ask questions such as these, or fail to ask them in a sufficiently basic and searching way.

Even though experiments have undeniable advantages (e.g. often, greater objectivity of experimental data; the use of controls to minimise the influence of extraneous variables; far greater power and severity for testing certain hypotheses), it does not follow that they are epistemically superior in every respect which has a bearing on the acquisition of knowledge in psychology and psychiatry. They may be positively disadvantageous (or counterproductive) in some respects[11], and/or lacking in some capacities which non-experimental ('informal') modes of inference possess. It therefore seems premature (and even methodologically illicit) to hand over epistemic control and authority wholesale to experimental methods without:

(a) Examining carefully the inferential and epistemic resources of 'ordinary (i.e. non-experimental) reasoning';

(b) Examining whether there are any respects in which such 'ordinary reasoning' (as applied in case-studies) might actually be advantageous over experimental (or controlled) studies;

(c) Carefully evaluating to what extent the generic problems and limitations of experimental methodology (both theoretical and practical) require us to make more modest claims on its behalf than if we had not taken those problems and limitations into account.

It cannot justifiably be assumed that because a study is an experimental one (i.e. has the hallmarks or trappings of an experiment) it will be able to provide a more discriminating and severe test of a given hypothesis (or will be able to establish the conclusion more securely) than an informal or non-experimental study. This applies (for some hypotheses) when experimental and informal modes of inference, respectively, are applied to the same hypothesis (e.g. it should not be assumed that an experimental test of a category II hypothesis h will necessarily be more effective in testing it and in securing a reliable result than an informal test of the same hypothesis). It also applies in an 'absolute' sense (e.g. it is possible that the particular informal test M of a category I hypothesis p is more severe, and yields a more secure conclusion, than the particular experimental test N does of category IV hypothesis q). To assume that experimental (including epidemiological) tests will always, will in every respect, or will necessarily be epistemically superior (e.g. more severe, more reliable, more conclusive) to non-experimental ones is an error[12] (we might call it 'The Fallacy of Automatic Experimental Privilege').

One may conjecture (somewhat speculatively) why some people fall victim to this fallacy. Here is my suggestion. Science in the modern world has demonstrated itself to be the most powerful and effective means of acquiring new and far-reaching knowledge. Experimental methodology is the primary – and perhaps distinctive – practical methodological tool of empirical science (Gower [1997] 10, 236; Hacking [1983] 149-150). Those wishing to participate in the epistemic success of science naturally infer that the best way to do this is to use and apply its practical methodology. Hence, conducting experiments is seen as engaging in methodological propriety. It may even come to be believed that the extent of the employment of experimentation is a measure of ones willingness to seek knowledge and truth (see e.g. the quotation from Eysenck and Wilson [1973] on page 22).

This reasoning is, in my view, overly simple. It does not appear to take adequate cognisance of the complexities of the way (or ways) in which we can and do acquire knowledge (see section 1.3). Distinguishing methodological propriety from impropriety in a way which corresponds, respectively (and strictly), to 'use of

experimentation' versus 'non-use of experimentation' (or use of non-experimental inference) is too crude. Moreover, the above reasoning does not appear to take adequate cognisance of what seem to be significant differences between the psychological (or human) sciences and the physical or natural sciences (see e.g. Rosenberg [1988] chapters 2 and 4). Whereas experimentation is obviously indispensable for physics, chemistry and many areas of biology, geology etc., it is not obvious that an experimental science of intentional mental states is feasible (though it would be unwise to rule it out dogmatically). On the other hand it would seem that we cannot conceptualise human psychology adequately or fully without the postulation and attribution of intentional mental states (I strongly believe this to be the case – see pages 43-44, 60, 69). Consequently, it may simply be that there is a substantial domain (i.e. intentional human psychology) for which an adequate experimental science cannot be developed, but of which concepts and hypotheses about that domain (and their being tested) is *essential* for psychological understanding. It may also be the case that informal (non-experimental) inference can provide a significant amount of reliable knowledge in this domain (e.g. knowledge of many of the mental and emotional states of oneself and others).

Overall, what I have wanted to oppose is the idea that experiments are always, or are in every respect, superior to non-experimental (i.e. informal) modes of inference (even when it comes to the testing of some causal hypotheses), and that this warrants transferring methodological authority to them, wholesale, even for psychology. It would, I believe, be more accurate to say that, depending upon the hypothesis and the exact nature of the surrounding circumstances, experiments (as they currently stand) sometimes are superior and sometimes are not.

If we consider Grünbaum's work I think we would have to stop short of saying that he hands over epistemic control and authority wholesale to experimentation. However, I think he comes very close to doing this. For example, although he says: "I do not maintain that any and all clinical [i.e. case-study] data are altogether irrelevant probatively", he adds "[but] this much only conditionally confers *potential* relevance on intraclinical results beyond their heuristic value" (Grünbaum [1984] 266; emphasis in original). He also says:

> "[O]n the whole, data from the couch *acquire* probative significance when they are independently corroborated by extraclinical findings, or when they are inductively consilient with such findings..."
> (Grünbaum ibid.; emphasis in original)

By saying "on the whole" in the last passage Grünbaum does implicitly allow for exceptions: for example, circumstances in which clinical inference could be sufficiently cogent to warrant acceptance of a conclusion without recourse to the authority of extraclinical (i.e. experimental or epidemiological) studies. But Grünbaum is reluctant to grant very much in this direction maintaining, at least for category III hypotheses (as opposed to category I and II ones), an overall scepticism towards experimentally unauthenticated conclusions. Whereas I agree with Grünbaum that extreme caution towards the clinical validation of many category III hypotheses is justified (at least, if the standard of validation we are asking for is very high), I am also critical about promoting experimentation as a potential remedy unless we can convincingly show that experimental studies can in fact provide the effective tests we desire. Without such a demonstration I can see little justification in implying that the reference standard of epistemic probity is experimentation (even though it might be in other domains, or for other hypotheses).

In spite of the above, I should emphasise that I am not in any principled way opposed to experimentation, but indeed am welcoming of it whenever it can be profitably and effectively applied (this is entirely consistent with holding an epistemologically critical attitude towards it). In my view, both experiments and (uncontrolled) case-studies are required by psychiatry as matters currently stand and as they are likely to be for the foreseeable future. Each have their epistemic strengths and weaknesses. The virtues of uncontrolled case-studies are, moreover, greater than that they should serve merely as forums for the generation of hypotheses in the Context of Discovery (with experiments serving to test those hypotheses in the Context of Justification). The latter is, however, the status to which

case-studies tend to be relegated by some critics. For example, Grünbaum says of Eysenck:

> "[I]n Eysenck's view, although clinical data from the couch may be heuristically fruitful by suggesting hypotheses, only suitably designed *experimental* studies can perform the *probative* role of *tests.*"
> (Grünbaum [1984] 97; emphases in original)

Grünbaum's own position does not strictly limit the role of psychoanalytic case-studies to one of hypothesis generation, but he does not credit them with much more than such a capacity (see e.g. Grünbaum [1980] 310; [1984] 189)[13]. In my view many uncontrolled case-studies have the capacity in their own right for allowing us to acquire valuable knowledge (including causal knowledge). This capacity may not be automatic or routine (being dependent upon the particularities of the case). Moreover, what can be achieved might be fairly modest. Nevertheless, I maintain, the capacity is present. This is something that is implicitly denied by those who maintain that they have *only* a hypothesis-generating ability and no effective testing and validational capabilities.

1.3 FP-REASONING AND EXPERIMENTATION

One way of proceeding in the philosophy of science is to focus on theories and their appraisal (see e.g. Popper [1959] 27, 59, 107-108; see also Hacking [1983] 149-150). The emphasis in this thesis will, however, be more generally epistemological. We can begin by considering what inductive modalities are available for potentially allowing us to acquire psychological knowledge. This approach is pertinent because the basic and most general question which has interested me is: "How do we acquire psychological knowledge of others or ourselves (if we do at all)?" There are certain dividends in going back to such a basic epistemological question rather than starting with some developed theory (such as Freud's) and evaluating it. One such advantage is that it allows us to look afresh at the complexity (and possible variety) of ways in which human beings learn or acquire knowledge at source. I have favoured such a 'naturalistic' epistemological approach[14] over a purely normative one – although I have also endeavoured never to let the normative methodological considerations slip very far behind. Normativism involves the stipulation of certain essentially *a priori* principles of rationality and the evaluation of whether actual methodological practice complies with these principles. On the other hand, what I mean by 'naturalism' implies giving priority to looking to see how we learn, how we acquire knowledge, what natural (in effect, biological or cognitive) capacities human beings possess which could possibly enable them to acquire knowledge, what is the nature of their practice when they attempt to acquire or extend knowledge etc. 'Naturalism', in the sense in which I intend it thus involves putting a form of empiricism (with regard to the question of how, as biological, cognitive and practical agents, we learn) ahead of methodological a priorism (or normativism). In spite of this, I also appreciate the indispensability of normativism and so regard my naturalism as modest and not radical. One consequence of such naturalism (and of an attendant pragmatism of which I shall say more in section 1.4(B)) is that in attempting to answer the preliminary question "How do human beings acquire psychological knowledge?", we need to distinguish two broad

inductive modalities both of which are potentially capable of providing psychological knowledge:

(i) (What I shall refer to as) Folk Psychological (or FP-) reasoning; and

(ii) Experimentation (which, as indicated on page 18, I shall terminologically take to include epidemiology unless a specific distinction between them needs to be made).

If we consider the range of available inductive modes and methods they include reasoning of the kind which is employed in everyday life and which does not make use of the special set-ups, controls, manipulations, ways of collecting data, or analytical techniques which are distinctive of scientific experimentation. We may call such reasoning 'ordinary reasoning' to distinguish it from the latter. The range of available inductive methods also includes, of course, just such experimental reasoning which is characteristic of developed science.

Ordinary reasoning includes – but is not exhausted by – FP-reasoning. (It also includes, at the very least, our everyday reasoning about the physical world or 'folk physics'.) There is some evidence to suggest that central aspects of FP-reasoning capacities are an ontogenetic outgrowth of normal biological/cognitive development (see e.g. Segal [1996] 150-156; Baron-Cohen and Swettenham [1996]; Botteril and Carruthers [1999] 52-56, 77), even if they can to some limited degree be augmented by a secondary (and conscious) application of normative reasoning skills (see page 41, subsection 2.2(iv), and section 3.2). Scientific experimentation is, by contrast, a cultural product (i.e. an invention of intellectual culture). Although there is some controversy as to when precisely concepts of controlled experimental methodology originated (Lindberg [1992] 358-361) the mainstream view is that they first appeared in anything like their modern form only with the Scientific Revolution of sixteenth and seventeenth century Europe (op. cit.). The idea of experimentation as a systematic way of enquiring about the world by manipulating it, or testing hypotheses about it, therefore seems to be an essentially modern one.

By FP-reasoning I shall mean the natural inferential and epistemic strategies and processes which *ex hypothesi* are employed when we attempt psychologically to comprehend one another in everyday life. FP-reasoning includes the reasoning which is involved in reaching conclusions about one's own mental and perceptual states and behaviour and their causes; as well as that which is involved in reaching conclusions of these kinds about others. At present we seem to lack an accurate and complete understanding of the precise inferential and epistemic features of FP-reasoning, as well as a satisfactory understanding of how it operates. Nevertheless, we must assume that inductively sound principles underlie it, if we accept that reliable FP-conclusions can be reached by its means. This is because it would be miraculous if reliable conclusions could be obtained without the utilisation of effective inductive principles and procedures – and we assume both that reliable FP-conclusions can be reached and that miracles do not occur. It is the (to a large extent tacit) use of such principles and procedures that I have in mind by the term 'FP-reasoning'.

I have accepted that insofar as FP-reasoning is effective it must rely on underlying principles of rationality for its operation. I think it is also the case that when we engage in FP-reasoning we sometimes engage in inferences which have certain similarities to (or are analogous to) canonical forms of reasoning identified by scientific methodologists (e.g. Mill's Methods for causal inference; Inference to the Best Explanation etc.). In spite of this, there are good reasons for viewing FP-reasoning as very unlike experimentation or experimental practice in the developed scientific sense. For example, unlike scientific experimentation, FP-reasoning does not make use of externally applied controls (even when causal hypotheses are at issue); nor does it normally involve manipulations of external physical conditions. Also, FP-reasoning operates naturally (and, it will be argued later, often effectively) with mentalistic postulates and concepts (such as intentions, emotions, beliefs etc.). On the other hand, scientific experimental reasoning (as it currently stands) is largely (or even completely) useless with regard to making independent discernments or discriminations and inferences about such mental states. (N.B. By "independent" I mean by the distinctive usage of its own resources, and not already making use of the

resources of FP-reasoning: it would hardly count that experimentation is making such discriminations or inferences if the actual work for that is being carried out by FP-reasoning.)

In distinguishing scientific experimentation from ordinary reasoning I am not claiming that when, in practice, scientists carry out experiments they abstain from engaging in ordinary reasoning or that, as an inductive procedure, experimentation could be conducted without a background reliance on ordinary reasoning. It is fairly obvious that experimenters make use of, for example, 'commonsense' (i.e. folk-physics) judgements about, say, the behaviour of materials or instruments (in addition to scientifically enlightened judgements about them); and that they also engage in FP-reasoning (e.g. to communicate with their fellow experimenters during the planning and execution of the experiment as well as when interpreting the result). None of this, however, detracts from the legitimacy of the claim that distinctive inductive contributions are being made by the different 'components' (i.e. broadly speaking, the experimental and ordinary reasoning components, respectively), and that these can, in principle, be distinguished. Indeed, it only goes to show that the totality of inferences involved in the practical conduct of an experiment involves a component which is due to the special experimental set-up, set of analytical techniques and/or distinctive experimental inference, plus a 'background' component due to ordinary reasoning. To fail to distinguish appropriately between the distinctive inductive contributions being made by each component leaves us in a welter of confusion about where our (possible) knowledge is coming from, how it is arrived at, and even whether we are entitled to it. According to the present account it would be mistaken to believe that when we move from relying on ordinary reasoning alone to using scientific experimentation we move from an unworthy (and/or defective) inductive modality to a wholly virtuous one. We have to keep using the (allegedly) inferior or unworthy inductive modality even as we engage in our experimental practice, and the strength of the experimental inference (and the warrant for its conclusion) never frees itself entirely from a dependence on 'lowly' ordinary reasoning.

As a general point of methodological significance, a distinction also needs to be made between the following:

> (i) A natural inductive process which is capable of delivering knowledge or information reliably (even though we may not know how it does this, and even though we may not have any manipulable control over its mode of operation). (N.B. such a process need not be perfect – it need only be effective and successful for a sufficiently large amount of the time.)
>
> (ii) Normative methodology. Normative methodology involves the practical exercise of certain rules or principles. We have a prior understanding that if these rules are followed certain types of inferential or epistemic errors will be avoided or certain inferential or epistemic gains made. This comes through a conscious understanding of how the rules work to circumvent the errors or to facilitate the positive gains.

An example of a (largely) reliable natural inductive process about which we had no adequate understanding and over which we had no control (until comparatively recently in human history) is normal human vision. Normal human vision could, by and large, deliver reliable knowledge about, say, the relative positions and sizes of macroscopic objects and their colours long before it was understood how it functioned (and there is still much that we do not understand about it – see e.g. Bernstein, Roy, Srull and Wickens [1991] 152). Moreover, prior to the development of ophthalmic lenses and eye surgery there was no significant scope for operationally manipulating this inductive capacity: the information or knowledge acquired by its means was the information or knowledge that the system (of eyes, brain etc.) delivered by virtue of its natural construction, internal functional capabilities, and performance in a certain range of environmental settings.

FP-reasoning is also, I suggest, essentially a natural inductive process (or set of such processes)[15]. It is not a method in the proper normative sense. There is only very limited scope for bringing the principles on which (we assume) it operates into conformity with our theoretical ideas about, for example, how to avoid inferential error, increase objectivity and severity etc.. Because the core principles of FP-reasoning are tacit (and, so to speak, built-in to the way our cognitive faculties function) there is effectively no flexibility for modifying them at all. There is scope for enhancing the inductive effectiveness and reliability of FP-reasoning at a secondary level, and this is something that all good psychotherapists would do. For example, one can be extra-attentive, patient and discerning as an observer (or listener); one can make as unbiased and judicious a use of the clinical evidence as is possible, and avoid deliberate suggestion; one can critically examine one's reasoning, consider alternative hypotheses, and be alert to potentially falsifying data etc.. However, all of these measures are structured on the judgmental capabilities of the psychotherapist. There are severe limitations in regard of what can be achieved by means of them. Nevertheless, I think it would be fair to say that they do introduce methodological normativity, even if only at a secondary level and in a very limited way.

Experimentation, on the other hand, is an example of the practical application of normative methodological principles. Scientific experiments are designed to eliminate or reduce certain types of inferential error, or to augment inductive capabilities through the implementation of special strategies, analytical techniques and/or practical set-ups. A central aim of experiments designed to test causal hypotheses is to reduce (hopefully eliminate) the influence of 'extraneous' variables, so that the influence of the presence/absence of the independent variable on the dependent one can be ascertained. There is, of course, considerable flexibility in how normative principles can be applied through the variety of different experimental designs.

The concept of validation is intrinsically normative. When we ask for a claim (hypothesis) to be validated it is not sufficient for it simply to be enunciated, even if it is true. Validation (and especially validation to scientific standards, or

scientific validation) requires at the very least: (i) a highly critical (and impartial) evaluation of the claim; and (ii) its demonstration (if there is indeed warrant for it). Contrast this with the yielding of a conclusion by a natural inductive process (e.g. by my vision, that the pen on my desk is farther away from me than the pencil; by my FP-reasoning, that Sue was angry with me when we met because I had forgotten to bring the tickets for our planned evening at the theatre). When a conclusion is yielded by a natural inductive process we often take it (usually unthinkingly) to constitute knowledge. I think it is not unreasonable to assume that natural inductive processes (such as vision, hearing and touch) often do yield reliable conclusions for which the term knowledge is apt. (A background concern for much of this thesis will be to what extent the same can be said for FP-reasoning.) However, it needs to be emphasised that such naturally yielded conclusions, even if reliable (or true) have not been validated (at least, not in an overt methodologically normative sense). It may be permissible to conceive of a given natural inductive process (such as vision) as carrying out its own intrinsic testing of a hypothesis and, let us suppose, validating it intrinsically, and then yielding this 'validated' conclusion as a result. But this would only be a *façon de parler*, an analogical or metaphorical use of the term 'validation' as compared to its proper methodological and normative sense. The latter requires openness to critical public scrutiny and objective demonstration to a very high standard.

From the foregoing it can be seen that a natural inductive process could issue in a reliable (or true) conclusion even though that conclusion has not been validated (and even if it could not be validated) in the proper sense. But although this is so, we clearly cannot automatically admit any statement which a given individual proclaims on the authority of his/her own natural inductive process(es). To do that would be to open the floodgates to all manner of bogus claims, since although natural inductive processes are (let us assume) often reliable, they may also err[16]. If the claim is of sufficient importance to us we would seek an independent and effective test of it, if that is possible (we should not assume that it always or necessarily will be). Experiments are methodologically important because what we hope to achieve through them are scientifically validated conclusions.

1.4 BACKGROUND THEMES TO THE THESIS

There are two themes which, whilst not constituting topics to be examined in their own right in this thesis, nevertheless constitute an important backdrop to it because they inform and guide many of the views expressed. Brief discussion of these themes is therefore in order:

(A) Mentalism

I am firmly committed to mentalism in psychology. By mentalism I mean a position which not merely accepts mental attributes (such as cognitive states, subjectively experienced perceptual and emotional states, and consciousness) as realities, but which proposes that a fully adequate account of human psychology and psychopathology cannot be given without taking such attributes to be centrally important (of no less importance than physiological and behavioural attributes). Mentalism does not, of course, imply anti-physicalism. Although I have no carefully worked-out position on the issue I have followed the majority of contemporary philosophers of mind (Kim [1996] 12) in accepting a materialist ontology or "physicalist framework" (Kim op. cit.).

Grünbaum is commendably sensitive to mental attributes in his analysis of Freud's work. This is so, for example, in his discussion of insight (see e.g. Grünbaum [1984] 130-140) and motivational explanation (op. cit. 69-83). However, I think that mentalism in the discussion of psychotherapy can and ought to go further. The concept of intentionality has a much wider scope than is covered by 'intending' (or an 'intention') in the motivational sense (Searle [1983] 3; [1994] 380). Indeed, perhaps the most interesting features of intentionality arise in connection with the analysis of certain attributes of mental states which have nothing specially to do with action or motivated behaviour at all: namely, the referential or representational 'aboutness' of many mental states and their (linguistically expressible) 'content' (see e.g. Kim [1996] 21). Whereas Grünbaum does discuss intentionality in relation to the explanation of motivated behaviour (op. cit.) there is, I believe, a lot more scope for making a relevant and profitable use of the concept in

analysing human psychology, including in contexts which are important for psychotherapy and psychiatry. We need a developed intentionalistic account of cognitive, emotional and even some perceptual states in order to have a satisfactory framework for describing the mental lives of others (or ourselves). For example, we need the intentionalistic concepts of 'aboutness' and/or 'content' in order adequately to describe or report such things as what a person believes (or is deluded about), what he/she fears (which may not even exist), is striving for (which might be unrealistic and unattainable), as well as a vast array of emotions, desires, memories, hopes, fantasies etc..

Treating mental states as intentional also raises some intriguing epistemological and methodological issues. One question which has interested me is: "What consequences does treating cognitive or emotional states as intentional have for how we can effectively test causal hypotheses about them?". Even though I do not examine this question in a systematic way in this thesis some issues raised by it are discussed in chapter 5 and sections 7.4 and 7.5.

(B) Pragmatic Realism with Regard to Hypothesis Testing

When philosophers of science discuss the testing of empirical hypotheses they often tend to emphasise the logical features of the test, so that the test can seem to be an exercise in applied logic (see e.g. Popper [1959] 32-34). Explicating the logic of tests and relating this to epistemic factors (such as observation or experience) is, of course, centrally important. Without a coherent inferential rationale a test would have no significance. However, there is another side to testing which is just as important but which often tends to receive less attention or can even be ignored altogether. This is the pragmatic (or practical) dimension of tests.

By the pragmatic dimension I mean all those factors connected with material conditions and with 'do-ability' (i.e. with what can or cannot be *done*) which in any way influences the performance or performability of the test. What information or knowledge can be gleaned from a test (and whether it can be conducted at all) depends on 'permitting' or 'restricting' material conditions and on

activities which can or cannot be carried out. Testing an hypothesis, as an epistemic operation, is thus ineliminably linked to what physical conditions can or cannot be brought into play, to what can or cannot be *done*. There is, I suggest, no test of an empirical hypothesis which is not an active, practical undertaking. The link between epistemology and some kind of active change in material conditions or circumstances is a profound one. This link is not evident if one focuses only upon the logical structure of the test (i.e. its inferential rationale). Without a suitable change in material conditions, whether through an external re-arrangement or manipulation of (some aspects of) the world and/or an internal physiological change (e.g. in one's sensory or cognitive apparatus) it would seem that no empirical hypothesis could be tested and no new knowledge concerning it could be arrived at. In a perfectly static universe in which there was no change it would seem that no hypothesis could be tested.

Pragmatic factors will influence – and may even govern completely – whether a test is possible in practice and, if it is, how severely or effectively the hypothesis can be tested. Because of practical or material constraints there may be no freedom to implement an otherwise straightforward testing logic.

There are at least two ways in which this focus on pragmatic factors is relevant for our purposes:

Firstly, any methodological advocacy of experimentation needs to take into consideration the 'implementability' of experiments. It is one thing to consider how controlled studies ameliorate or overcome the inferential liabilities of uncontrolled studies when all that is being given is a theoretical analysis of those virtues. It is quite another matter as to whether a well-controlled study can actually be set up and effectively applied (in practice) to test the hypothesis at issue. When it comes to implementation in practice – which is the only circumstance that ultimately matters – an experiment may fall well short of being able to instantiate those normative principles which were the (theoretical) reason for its being selected, or it may fail completely.

Secondly, although experimental tests in principle exemplify methodological virtues much more than informal (uncontrolled) inference, if no

experiment can in practice be applied to test a given hypothesis, but an informal test (i.e. informal inference bearing on the hypothesis) can, in practice, be carried out then, so long as the informal inference can be judged to be of an acceptably high degree of severity and reliability, it may well be preferable to experimentation. In a situation like this informal inference turns out to be more effective (and preferable) to experimentation primarily for pragmatic reasons. A testing format which exemplifies our desire to follow methodologically normative ideals but which cannot be applied in practice is of no use at all. A format which (we acknowledge) falls well short of methodological ideals but can be applied in practice (and if we have good reasons to suppose it is effective a lot of the time) may be adequate for our epistemic needs. FP-reasoning, I believe, has just this kind of pragmatic advantage over experimentation for the testing of many hypotheses in the psychological domain (especially in categories I and II).

CHAPTER 2

THE IMPORTANCE OF FP-REASONING

AND SOME OF ITS LIMITATIONS

2.1 FP-REASONING IS IN SOME RESPECTS INDUCTIVELY SUPERIOR TO EXPERIMENTATION

Is it ever the case that the inferential and/or epistemic capacities of FP-reasoning – as, for example, might be employed in uncontrolled case-studies, or even in daily life – are inductively superior to those of experimental or epidemiological methods? That is, are there any circumstances in which FP-reasoning unaided by any of the distinctive arrangements or techniques of experimental and epidemiological methods can *outperform* the latter; either in terms of the kinds of hypotheses which can be tested (or validated) by its means, or in terms of the adequacy with which the test can be carried out (e.g. regarding the severity of the test or the reliability of the conclusion)? If the answers to these questions were affirmative then there would be no justification for regarding experimental studies in psychology as being always or in every respect inductively superior to FP-reasoning, and there could be no justification for an automatic methodological prescription in their favour. Mickey Mouse's favourite food is cheese.

Let us examine these questions by considering some examples; although it may help if I state at the outset that my own belief is that they can be answered affirmatively.

Consider, first, the hypothesis that you (the reader) has, a short time ago, thought about what is semantically represented by the syntax of the sentence 'Mickey Mouse's favourite food is cheese'[17]. I reach the conclusion that this hypothesis is probably true on the basis of inferences and tacit theory that is typical of FP (even though it would be difficult or impossible to reconstruct the details fully). It involves background assumptions (or theory) about the likely effect on your mentation of certain symbols which I hypothesise you perceive and cognitively interpret. The hypothesis is important because it potentially tells me something about your psychology (mental states). As a hypothesis it is fallible. It is possible that you have not recently (say, within the last two minutes) had such a thought. For example, your concentration might have lapsed while reading this text; or you might have

started reading this paragraph after having skipped the penultimate one. But if you did read the penultimate paragraph a short while ago then I judge that it is very likely that you have recently had such a thought (am I correct?)

There are two points of importance:

First, if one had the opportunity of talking to you, the reader, it would most likely be possible to ascertain with a good degree of reliability and accuracy whether the hypothesis was true or not. That is, it would be possible to test/validate the hypothesis (by the informal, not scientific, means that FP-reasoning provides), yielding a fairly reliable conclusion. The inference would, of course, also make use of the assumption that you are capable of veridically recognising certain of your own mental states, as well as reporting them reliably – and that you will do the latter truthfully.

Secondly, there is, to the best of my knowledge, no (extant) independent experimental or epidemiological method by which the hypothesis can be effectively tested/validated. By an "independent" method I mean one which does not already make primary use of the inferential and epistemic resources of FP-reasoning to test the hypothesis, but accomplishes this by means of its own distinctive resources. If it is in fact the case that the hypothesis at issue was tested to an acceptable degree of reliability using FP-reasoning, and that this was not achievable by means of any extant experimental method then, for this example at least, FP-reasoning turns out to be more inductively powerful than any extant experimental method.

Let us now turn to some examples that are, potentially, clinically relevant. There is an indefinitely large set of descriptive (as opposed to causal) statements which psychotherapists would be interested in which describe aspects of their clients' psychologies in terms of intentionally directed mental (i.e. cognitive or emotional) states. These can be conceived and formulated as hypotheses, in the sense that in attempting to ascertain the nature of the mental state and its intentional content one uses evidence (for example, behavioural or verbal) to support a conclusion about it. For example:

Hypothesis h_1: Jane harbours a 'secret' fear of being left all alone in the world, without family or friends[18].

Hypothesis h_2: Jeff is obsessively fearful of encountering aggressive, anti-social or criminal behaviour from members of the public.

Hypothesis h_3: Steve is preoccupied with feelings of disgust at the sexual activities of homosexuals.

Hypothesis h_4: Jill is preoccupied with feelings of disgust at the sexual activities of her parents.

Hypothesis h_5: Mary feels contempt for all women who marry and have children.

Hypothesis h_6: Tom feels guilt for having 'rejected' or 'disowned' his father. (This example is taken from the case-study dealt with in chapter 3.)

It is an essential part of the work of psychotherapists to ascertain accurately leading features of the emotional and cognitive lives of their clients. I think it can justifiably be maintained that FP-reasoning offers a reasonably – and sometimes very – effective means of doing this for an important range of cognitive and emotional states[19]. For example, by talking to Jane (hypothesis h_1, above), listening to her patiently and sympathetically, asking her appropriate questions, encouraging her to articulate her feelings etc. we may be able to learn with a high degree of reliability that she harbours a fear of being alone, without family or friends. Evidence which came up which was contrary to the hypothesis would, on the other hand, require us (provisionally) to reject it. It might, moreover, be possible to cross-check our conclusion with evidence bearing on the hypothesis obtained in different contexts of discussion on separate occasions. All of the above applies equally to $h_2 - h_6$ and to indefinitely many other hypotheses covering a wide range of emotional and cognitive states.

Compare the latter with the ability of extant experimental or epidemiological methods to test the same hypotheses. Are there any effective independent experimental tests of $h_1 - h_6$, or comparable hypotheses? The answer seems to be "no". If this is the case then, insofar as use of FP-reasoning can yield reliable conclusions, it provides a testing and validational capability which current

experimental methods cannot equal. I do not think this is a trivial accomplishment. If the argument is sound it implies that there are important features of Jane's psychology (also of Jeff's, Steve's etc.) that we cannot learn about whilst relying on the distinctive methodological devices of mature science, but which we can often learn about (with a good degree of reliability) by using FP-reasoning.

Does the above verdict apply also to causal hypotheses? That is, are there any respects in which FP-reasoning can outperform the mature experimental methods of science when it comes to testing causal hypotheses? It certainly seems to be the case that there are some circumstances in which, by relying on FP-reasoning plus a substantial amount of background knowledge (of a kind that is typically acquired through 'commonsense' reasoning and experience as opposed to scientifically tested and validated theories) a highly reliable causal conclusion can be reached which no extant experimental method could have effectively tested or validated. An example is the hypothesis (and conclusion) T1 in my 'Tom' case-study (the reader is referred to section 3.34). The same verdict probably also applies to hypotheses T2, T3 and T4A from the same case-study although, in the case of T3 (and possibly T4A too), the specific motivational nature of the hypothesis may make the grounds for inferring a causal connection more problematic (see pages 93-94; see also 96-98). There is, moreover, no good reason for supposing that T1-T4A are isolated examples.

The point of making this comparison between FP-reasoning and experimentation is not to initiate a competition to demonstrate which is 'better', overall: that would be silly. Obviously, there are many hypotheses which, if they are going to be tested effectively at all will require the use of mature experimental methods and practices. The point is, rather, that circumstances are not so simple or uniform over all domains in which we seek knowledge that we can expect that mature experimental methods can be routinely (or automatically) applied to provide the most severe or effective test or the most reliable result. Informal methods (such as making primary utilisation of FP-reasoning) are sometimes better: it depends on what the hypothesis is plus a host of other, including circumstantial, factors.

Regarding the severity of a test (i.e. its ability to demonstrate the hypothesis to be false, if it is so), if an experimental format cannot be applied in practice then, by default it provides no severity at all. An informal test of the same hypothesis (e.g. h_1 or T1, mentioned above), though not ideal, may nevertheless provide a substantial degree of severity and might, on the basis of the available evidence, command a sufficient degree of rational warrant to make the conclusion 'beyond reasonable doubt'.

Regarding the degree of certainty with which a conclusion can be established, even well-conducted experiments can yield conclusions which are inconclusive (e.g. Chassan [1979] 210). On the other hand, many conclusions which are yielded by informal or FP-reasoning can be deserving of a high degree of rational credence (e.g. my 'Mickey Mouse' example; and hypothesis T1). Consequently, there can be no guarantee that an experimental study will establish a conclusion with greater certainty than an informal inference based on FP-reasoning.

2.2 A COMPARISON BETWEEN FP-REASONING AND EXPERIMENTATION

We shall now compare the inductive capabilities of experimentation with those of FP-reasoning across a variety of epistemic and inferential criteria. Obviously, caution is needed in making a comparison at such an abstract level. Nevertheless, it is felt that the differences outlined are broadly accurate generalisations. The criteria are: (i) objectivity; (ii) inferential explicitness; (iii) scope for eliminating inferential error; (iv) severity.

(i) Objectivity

The data used in experimental and epidemiological studies are, in general, more objective than those which are employed in FP-reasoning. By this, I mean both that there will be greater scope for the intersubjective accessibility of the data, and that there will be definite rules (agreed upon in advance) for enabling us to decide upon a given datum (e.g. what constitutes a measurement; that a given physical state obtains; that a given physical transformation has taken place etc.). The data used in FP-reasoning may not be altogether lacking in this kind of objectivity (although measurement is not normally involved), but they will in general be more impoverished with regard to it, and often very much so.

By relying on behavioural data (which, of course, is intersubjectively accessible) we may be able to agree fairly objectively that Fred is much less active than usual (he looks lethargic, stoops a little, does not engage in his usual interests etc.). However, reaching behaviouristic conclusions about someone is not what FP is primarily about. FP is paradigmatically mentalistic (Wellman [1990] 2, 4-5, 98; Baron-Cohen [1995] chapter 1), and when we make inferences about Fred's mental states, motives etc. the evidence we use to reach our conclusions may fail to satisfy strong conditions of objectivity. I may interpret Fred's condition as stemming from depression, you may interpret it as being due to his being fatigued (but not depressed). When we attempt to reach conclusions about mental states we increase the likelihood of utilising data that are not as objective as the crucial data used in experimental and epidemiological studies. Also, we often use FP-reasoning to reach

conclusions about our own mental states or motives (for example, that one is depressed and not merely fatigued; feels guilty about p; is angry about q etc.). In this case the data utilised in the inference may be grossly non-objective (e.g. one's subjective feelings).

In spite of a general weakening of the objectivity of data when making mentalistic (as opposed to behavioural) inferences, there is one facility which can be exploited to increase dramatically their objectivity as well as the accuracy of the conclusions reached. This is what James Hopkins calls the 'linguistic articulation' (or articulability) of mental states (Hopkins [1991] 88-89). Philosophers of mind have long recognised that intentional mental states (such as beliefs, desires, and many emotional states) have an intimate (if not fully understood) relation to semantically specifiable propositional clauses (see e.g. Perry [1994]). The latter can be used as descriptive 'markers' for a wide range of cognitive, motivational and emotional states. If one wants to know whether Fred is depressed or just fatigued it would assist our attempt to reach a reliable conclusion by talking to him and eliciting a report from him about how he feels. This attempt may be inconclusive (or perhaps he is both depressed and fatigued); and, as with all inductive inferences it is, of course, fallible. The point is, however, that by utilising linguistic data relating to the intentional content of his mental states we gain a considerable epistemic advantage in respect of being able to use objective data. What Fred says (or even writes down) is pretty objective as data. Moreover, on the basis of certain theoretical assumptions we can take the linguistic articulations as reasonably reliable indicators of the content of his mental states (thereby acquiring inferred knowledge about them). For example, if Fred says: "I've been feeling very unhappy – my life feels empty and without purpose", we can – *ceteris paribus* – infer that this indicates his subjective cognitive/affective state. On background knowledge it is likely that he is depressed.

(ii) Inferential Explicitness

The plan of the key inference is usually quite explicit in experiments designed to test specific hypotheses. (Note: Not all experimental studies will be designed to test specific hypotheses. Some will be exploratory – carried out just to

see 'what happens'.) For example, in a randomized clinical trial (RCT) or a standard single subject ABAB design (see section 4.2) the core inferential strategy by which the hypothesis is tested can be articulated and is unambiguous. In FP-reasoning the crucial inference is often less explicit and less articulable. For example, I might reach the conclusion (even quite accurately) that Bill's jealousy of Tom's success motivated (caused) him to try to undermine Tom's enterprises. However, it may be difficult or impossible to lay bare the exact steps in this inference, or the crucial ones by which the hypothesis was tested. It is not being claimed that FP-inferences are altogether obscure. We can often piece together various lines of reasoning (e.g. some variant of the practical syllogism) by which a conclusion can be justified or a hypothesis (informally) tested. But, unlike the inferences in the aforementioned experiments, these are typically not made explicit beforehand and then subsequently utilised. Rather, FP-reasoning usually proceeds at more of an intuitive level; and any account of the steps used are then subsequently reconstructed, if one is called upon to provide it. This is less likely to issue in an accurate account of the actual inference used than if the steps were laid out in advance and then followed. Moreover, the inferences actually used in reaching FP conclusions by FP-reasoning seem in general to be less amenable to being objectively analysed and displayed (than are the key inferences in experiments).

(iii) Scope for Eliminating Inferential Error

Experiments have, in general, far greater scope for the incorporation into their inferences of strategies or devices capable of reducing certain kinds of inferential error. This is so especially in the case of causal inferences. Experimental controls can be introduced to block the influence of variables whose influence are known about but are undesired. Other strategies - such as randomization - can be used to 'equalise' the effects of extraneous variables whose individual effects cannot be controlled. (The effectiveness of these strategies need, however, to be considered critically - see chapter 4.) Often, the effectiveness of the controlling or randomizing strategy can be assessed as part of a separate investigation, prior to its incorporation into the main experimental or epidemiological study at hand. There is really nothing comparable to

this in the case of FP-reasoning. With FP-reasoning, once one has made an inference to a conclusion on the basis of various data, one can reconsider whether a particular piece of data is as reliable, or the inference is as cogent, as one at first thought. One might then alter one's judgement on the matter. However:

(a) There is no significant scope within the inferential processes constitutive of FP-reasoning for the interposition of measures comparable to experimental controls and randomization;

(b) Attempted improvements in inferential soundness consist principally of attempted improvements of one's judgement on the matter. Within the rubric of FP-reasoning there is little or no scope for eliminating error by means which are independent of such judgement.

(iv) Severity

The potential severity of experimental tests is, in general, much greater than that of FP-reasoning. We should, however, note that there will be considerable variation between different studies (some studies may not be very severe). By severity we mean the capacity which the test has for showing the hypothesis to be false, if it is so. The severity which experiments (and epidemiological studies) possess derives from a combination of:

(a) The inherent logic of potentially falsifying the test hypothesis which is built into the design; and

(b) The kinds of auxiliary strategies which can be added (such as the controlling or randomizing features already mentioned). These can be used to assist in focusing the crucial test inference more precisely by minimising the influence of extraneous variables.

In FP-reasoning, as we have seen, there is really no scope for introducing (b). There may be some limited scope for increasing the severity of a judgement. For example one might:

(1) Enlarge the potentially falsifying data-base one is using; or
(2) Adopt a more critical attitude towards a hypothesis one initially gave credence to; or
(3) Be more explicit and critical about the lines of reasoning by which one reached a particular conclusion, or about the evidence on the basis of which one reached it.

However, these measures – with the possible exception of (1) – are informal and are based on judgement, possibly of an intuitive nature.

Where FP-reasoning does, I believe, have some advantage over experimentation with regard to the severity of the tests it can carry out, is when experiments (or epidemiological studies) cannot be applied in practice to test the hypotheses in question at all. This is the scenario we considered on pages 45-46. In this case experiments, by default, have no severity whatsoever. If, on the other hand, FP-reasoning can provide a reasonably effective test (even if it is a comparatively weak and informal one) of at least some of these hypotheses then, in practice, it will be more severe with regard to them.

FP-reasoning also has at its disposal the means of a quasi-increase-in-severity by exploiting the facility of the linguistic articulation of mental states mentioned earlier (subsection 2.2(i)). By posing more and more specific and searching questions of an appropriate kind (and assuming that the responses are truthful, and are delivered by an epistemically reliable process) we can conduct more and more severe (albeit informal) tests on a range of FP or FP-like hypotheses. For example, if we want to 'test' whether Bill is jealous of Tom, whether Jeff has suicidal inclinations, or that Sue's disgust is about p, we can engage in a dialogue with the respective individuals in order to evoke responses which bear upon the truth or falsity of the corresponding hypotheses in a progressively severe way.

The foregoing way of increasing the severity of the test has been described as 'quasi-'. This is because it relies on semantic content (that of the questions and the answers) to extract information about the subject's psychological states. *Ex hypothesi*, there will be many cognitive, emotional or motivational states of an individual which are (or can be) linguistically articulated by specific propositional clauses. These states and their articulations will initially be unknown to an external observer. On the assumption that the subject is capable of veridically recognising those states (through introspection) and can reliably report them to us, we can put to him/her questions which verbally test our hypotheses about those states. For example, we may hypothesise that Bill is jealous of Tom's success. We could begin by asking Bill whether he admires Tom's achievements. Suppose he answers affirmatively. We could continue by asking Bill whether he feels he has achieved less than Tom. If his answer is again affirmative, we could ask whether he is envious of Tom's achievements. We could then proceed by asking whether he feels any bitterness or resentment towards Tom because of Tom's greater success etc.. In a procedure such as this we semantically tighten the constraints on the possibility of the hypothesis being shown to be either true or false and, in this sense, increase the severity of the test on it.

The ability of the above question and answer procedure to provide an increasingly severe (informal) test of the hypothesis at issue is, of course, predicated upon a variety of inferential (including evidence-utilising) and cognitive processes operating reliably 'in the background'. For example, it depends upon the subject being able to understand the questions put to him/her; to apply these in a relevant and critical way to his/her own mental or motivational states; to recognise the 'fact-of-the-matter' with regard to his/her own cognitions, emotions or motivations; and to make truthful reports about the latter, sometimes in a progressively accurate and complete way. What the question and answer process takes for granted, however, is the independent justification of the steps which would guarantee that it is a reliable epistemic process. Yet it is the latter that the scientific sceptic asks for. The standard scientific conception of a test would not allow it to be based on a verbal dialogue of

this kind (together with all the inherent assumptions about the epistemic reliability of the processes involved)[20].

2.3 THE IMPORTANCE OF FP-REASONING FOR MENTALISTIC UNDERSTANDING

The reason for valuing and advocating (experimentally uncontrolled) FP-reasoning does not stem from a perversity to promote what is scientifically inadequate. It has already been acknowledged that FP-reasoning is in many respects inferior to testing and validation by experimental and epidemiological methods (see section 2.2). The advocacy of FP-reasoning stems, rather, from two related considerations the importance of which are felt to outweigh its limitations and liabilities.

Firstly, I regard it as axiomatic that psychiatry should possess the means of understanding others (or oneself) in mentalistic terms. A mentalistic understanding – in which intentional mental states are posited for descriptive and explanatory purposes – is simply too fundamental to psychological (and psychopathological) understanding to be excluded or compromised. A psychiatry which does compromise on this is simply a psychiatry not worth having (whatever else it may achieve by way of, say, behaviouristic or neurophysiological understanding).

As matters currently stand, the best (and, to a large degree, the only) mentalistic psychology that we have that can be applied for descriptive, explanatory and (to a limited extent) predictive purposes in real-life settings is FP (Baron-Cohen [1995] 21-26; Bolton and Hill [1996] 30-32; Dennett [1987] 21; Fodor [1987] x, xii, 9-10). In other words, if we want to describe, explain or predict human action and mentation in real-life social and worldly settings we have little or no option but to use the postulates and conceptual framework of FP.

Secondly, regarding a great many of the hypotheses in category I (i.e. FP-hypotheses) but also perhaps many in category II (i.e. FP-like psychotherapeutic hypotheses), there is no other way to test/validate them than by relying on FP-reasoning. I do not rule out the possibility that some of these hypotheses will eventually be effectively testable by experimental means (especially if they are generic claims). However, even if this is the case there may still be many which

cannot be tested at all except by FP-reasoning. In that case our choice comes down to not having these hypotheses tested/validated at all, or having them tested/validated by the (scientifically inadequate) inductive modality of FP-reasoning[21]. Faced with these options (and given my uncompromising commitment to mentalistic understanding) I think it is preferable to rely on the less than entirely satisfactory modality of FP-reasoning.

Concerns about the unreliability of FP-reasoning can also be exaggerated. A strong and systematic scepticism towards the attainment of FP knowledge (given that we admit the ontology of FP) is, I believe, untenable (I argue for this in 7.22). However, there remain genuine and major difficulties with regard to the provision of adequate justification for FP-claims. I do not pretend to be able to provide a solution to these in this thesis. We need, at the very least, to draw a distinction between:

> (i) A *global* justification of FP-knowledge. This is an attempt to counter the sceptic who maintains that we have no rational basis for believing that we acquire reliable FP-knowledge (including causal FP-knowledge) on a fairly regular basis in ordinary social life; and
>
> (ii) Our ability to test/validate *individual* FP (or FP-like) hypotheses (be they singular or generic) in conformity with epistemic standards that are typical of scientific demonstrations at their best.

Whereas I think '(i)' can be accomplished (see 7.22), it is likely that '(ii)' cannot be for a large number of (even if not for all) FP or FP-like claims. It would appear that it is simply not possible, without moving to a set of demonstrative resources which lie outside of FP-reasoning (and which FP-reasoning lacks), to provide a demonstration of an FP-claim to the standards referred to in (ii). And as we have seen, when we move outside the resources of FP-reasoning there are many FP-hypotheses which we may not be able to test at all.

2.4 SCIENTIFIC VALIDATION IS NEITHER NECESSARY NOR SUFFICIENT FOR THE RECOGNITION OF THE TRUTH OF SOME NON-TRIVIAL PSYCHOLOGICAL CLAIMS

Scientific validation is a very demanding mode of justification. It requires the satisfaction of a variety of epistemic criteria either to standards that are much higher than would be required outside science, or that would not be required to be satisfied at all outside science. The epistemic criteria include: objectivity of data; explicitness of inference; the severity of the test; the degree of certainty of the result. Many critics ask for the scientific validation of any non-trivial empirical claim which aspires to receive full rational acceptance, and this would include the claims of psychoanalysts and psychotherapists.

I shall argue below that scientific validation is neither necessary nor sufficient for the recognition of the truth of some non-trivial empirical truths of psychology. My aim here is merely to show that this is logically and epistemically possible. I shall not attempt to argue for the further claim that there are, *in fact*, some very important empirical truths of psychological interest which fall into this category (although this could be done). My claim is important in light of the fact that (to simplify somewhat) the central message of Grünbaum's critique of psychoanalysis is that scientific validation should stand as the criterion of hypothesis-acceptance, at least for category III hypotheses. It should become clear from the following discussion that: (i) there could be certain (non-trivial) psychological truths that could be known reliably, yet independently of their being scientifically validated; and (ii) the totality of scientifically-validated hypotheses (including those yet to be so validated) could permanently exclude some personally knowable (non-trivial) psychological truths. If correct, this has important implications for psychological - including psychiatric - methodology. Scientific validation would fail as an epistemic principle for the disclosure of non-trivial psychological truths; yet subjective knowledge of such truths could not serve as an epistemically sound basis for their general acceptance.

Let us begin with a trivial example. Consider the following statement, p:

> p: "On a day in August 1983, whilst walking alone along the coastal path between Budleigh Salterton and Sidmouth (in South Devon), I (Michael Dash) imagined the surrounding landscape as it might have been at a remote period of the Earth's history."

During a holiday in Devon in August 1983 I had been walking along the coastal path as stated. I had been thinking about evolution and the age of the Earth and tried to imagine the landscape as it might have been at a remote time with Plesiosaurs and Ichthyosaurs swimming in the sea.

p, I maintain, is a true empirical claim. p says something about what I imagined and also when and where. I accept that it is not ideally precise (I could make it a little more precise). But I believe it is sufficiently precise to have a truth-value. p makes an assertion about what I imagined and, thus, about my psychology. It would have been possible to have used other real-life examples about other aspects of my psychology at various times - for example, about what I have experienced, feared, hoped for etc..

I have asserted that p is true (and it is!), but how can you be sure? It seems that you only have my word for it and, if you adopt a sceptical attitude, you may doubt the reliability of my memory, or even treat my claim as a deliberate hoax.

It might be felt that in order to be confident about whether p is deserving of rational credence or not we could do no better than attempt scientifically to validate it. Scientific validation, after all, provides the justificatory standard of that corpus of empirical knowledge (which is not of an elementary perceptual nature) which is widely regarded as the most certain and important that we possess – namely, science.

But, I maintain, p cannot be scientifically validated. There is no independent scientific test (or set of tests) that can be applied effectively to test it. Or, at least, there is none that does not already make use of my first-person

introspective capacity to know whether p is true – and the latter, of course, does not constitute the scientific validation of p. (If the reader can provide a counterexample to my claim I am prepared to stand corrected[22].)

On the other hand, if you grant me that I am not hoaxing and that my memory is reliable you may well believe that p is true. But even if you do not, it is still logically possible that p is true and that I know it is. If p is true, the possibility of my recognising it as true (as mediated by a reliable process of memory) does not depend on whether you believe it to be true or not, and neither does it require the scientific validation of p. It is logically possible that I can know p to be true (if it is), even though p cannot be scientifically validated.

This is not a trivial point. On a standard Realist interpretation of knowledge and science (according to which science discloses the truth) we would not, perhaps, have expected that scientific validation and the recognition of what is true could have been so seriously out of synchrony. What can be subjectively recognised as true (if it is) as part of one's personal or self-knowledge may not be capable of being scientifically validated. However, as I shall point out below, into this category may fall various hypotheses (or claims) which are psychologically important, both theoretically and practically.

So much for the argument concerning necessity. Let us now turn to the question of whether scientific validation is sufficient for the recognition of the truth of p (by a single agent).

Suppose that the only empirical statements that we were allowed to accept (which were not themselves elementary perceptual statements) were those that had been scientifically validated. Suppose also that there are effective procedures (such as experiments) for testing/validating such (non-elementary) empirical hypotheses to scientific standards. This would include procedures currently available and also, possibly, some new ones in the future. Does it follow that we could be absolutely sure that p could and would be scientifically validated if we operated these procedures for an indefinite length of time? In my view, no. To the best of my knowledge there is no extant set of procedures which could scientifically

validate p. And even if some new methods of validation became available in the future we could not be sure that they could serve that end. Indeed, once I have died and my brain has been dissipated it would seem that any attempt to recover information in objective scientific terms about what I imagined in my life will be impossible. It is not logically impossible that p could be scientifically validated, but the likelihood of this is so remote as to be virtually zero. Consequently, even if we ran effective procedures for scientifically validating hypotheses for an indefinite length of time we may still never be able to validate p to the appropriate standards. This is the case even though, if my memory processes are reliable, they would guarantee my knowledge of the truth of p, if p is true (which it is). Scientific validation is not sufficient for the recognition of the truth of at least some true empirical hypotheses.

Whereas p is of no significant psychological interest, it is possible that there are true claims of comparable epistemic standing which *are* important. This might be because the experiences or events recalled in them are of personal (including traumatic) significance to the individuals concerned, and/or because the experiences or events are theoretically important for psychology. Suppose, for example, that Sue (who is now an adult) is capable of recognising the truth of q, where q is the true statement:

> q: "When I (Sue) was about ten years old I was repeatedly sexually abused by [a certain person]."
> (We could even assume that there are no witnesses able or willing to confirm that Sue was abused.)

It is also possible that there are distinctively Freudian examples. For example, suppose Harry recognises the truth of r, where r is the true statement:

> r: "As a small boy, after seeing my sister in the bath, I (Harry) began to think that she had once had a penis and began to fear that mine would be removed too."

The latter might constitute strong evidence for what Freud called 'castration anxiety', especially if Harry's anxiety was self-generated and did not arise in response to any external threat or provocation.

An essential requirement of scientific validation is the public accessibility of the data which is crucial for the inductive justification. This is lacking in the case of first-person testimony. But note that lack of public accessibility of data does not entail the falsehood of a conclusion based on that data. Data may be publicly inaccessible (and intrasubjective) and yet a conclusion based on that data may be extremely reliable or true, if the processes involved in generating the conclusion from the data are themselves truth-preserving.

2.5 AN HISTORICAL EXCURSUS – FREUD'S DISMISSAL OF EXPERIMENTATION

One interesting corollary of my defence of mentalism in sections 1.4(A) and 2.3 (despite its adverse methodological consequences) concerns the historical interpretation of Freud. Various commentators have drawn attention to a celebrated postcard sent by Freud to the experimental psychologist Saul Rosenzweig (Freud [1934]). Rosenzweig claimed to have tested Freud's theory of repression experimentally – with a favourable outcome – and notified Freud of the result. (In fact, it seems that the type of repression Rosenzweig tested was not distinctively psychoanalytical, but a 'weaker' cognitive version – see Grünbaum [1984] 102. This misconstrual is, however, irrelevant for present purposes.)

Freud then replied to Rosenzweig as follows:

> "I have examined your experimental studies for the verification of the psychoanalytic assertions with interest. I cannot put much value on these confirmations because the wealth of reliable observations on which these assertions rest make them independent of experimental verification. Still, it can do no harm."
>
> (Freud [1934]. Quoted in each of Grünbaum [1984] 101, Glymour [1974] 286, Eysenck [1985] 149-150, and Eysenck and Wilson [1973] xi)

We cannot, of course, accept at face value Freud's assurances that his conclusions based on clinical observations are sound. In this thesis I have been both critical of FP-reasoning and attempted to give recognition to its virtues. Also, as Grünbaum points out (Grünbaum [1993] 112), clinical interview testing becomes more suspect as we proceed from category I and II hypotheses to category III (which is the category into which the above Freudian claims would fall). Nevertheless, the general question of why Freud was almost completely dismissive of the experimental testing of psychoanalytic claims (see also Freud [1933] 210-211), and yet favourably disposed toward clinical testing (which is fundamentally dependent upon FP-reasoning), remains intriguing. A standard view is that it is because of Freud's incompetence as a scientific methodologist. Eysenck, for example, says:

> "[Freud's] attitude to . . . the experimental approach . . . [, which is] probably the most decisive and convincing scientific method, is revealed in his famous postcard to Rosenzweig . . . . Nothing could demonstrate more clearly the non-scientific character of Freud's thinking; in his view, experiments were not needed to confirm his hypotheses, nor could they influence them."
>
> (Eysenck [1985] 149-150)

However, any attempt to portray Freud as a scientific methodologist of poor quality is highly implausible. There is no denying his eschewing of experimentation, as we have seen. However, this would necessarily count against his quality as a scientific methodologist (and epistemologist) only on the overly simple (and erroneous) view that good scientific practice is always conditional upon performing experiments. There can be no doubting serious lapses and blind-spots in Freud's methodological and epistemological competence if we take his corpus as a whole (see e.g. Glymour [1980] 263-265). However, Grünbaum judges him to be a "sophisticated scientific methodologist" (Grünbaum [1984] 128) and says that "Freud obviously had a keen appreciation of the methodological safeguards afforded by controlled prospective causal inquiry, no less than of the pitfalls of *post hoc ergo propter hoc* inferences" (op. cit. 170). In my own estimation Freud is at least as competent a scientific methodologist and epistemologist as Eysenck. That level of competence is, moreover, good – especially if we appreciate that they are professional psychologists, and not professional methodologists.

I now wish to present a different interpretation of Freud's favouring of the clinical interview method over experimentation. This is in keeping with the ideas expressed in section 2.3.

I suggest that Freud favoured FP-reasoning over experimentation not necessarily because he believed it to be superior in the abstract for the kinds of epistemic criteria we considered earlier (see section 2.2). Rather, he preferred it because he valued mentalistic understanding. He regarded the latter as a theoretical prerequisite for the adequate understanding of human psychology and psychopathology, just as I have done. FP-reasoning (and the clinical approach) then became the inductively expedient means of attaining such mentalistic understanding.

In spite of its weaknesses it was still more effective in achieving this end than reliance on experimentation (as an autonomous method) would have been.

There can be no doubt that Freud was a Mentalist and that Freudian psychoanalysis is a mentalistic theoretical system. (Indeed, all psychodynamic psychological systems are mentalistic.) Evidence for this can be found throughout Freud's writings: Freudian explanations are typically in terms of the mental states of the subject. For the sake of brevity the reader is referred to (Freud [1924] 161-162 and [1925] 265-266) where Freud points out that it is precisely the "psychical" (i.e. mental) factor (op. cit.) which psychoanalysis attempted to take account of, and which distinguished it from the orthodox psychiatry of the day (which did not). (See also Freud [1916] 129 and [1917](b) 319, 321.)

It is only by sidelining or eschewing mentalism that one can, coherently, treat experimentation as a sufficient practical methodology for psychology. This may not matter to a Behaviourist (which is what Eysenck is) or a neurophysiologist. But then there is no current reason to suppose that these psychological research programmes will ever be adequate for explaining human psychology. Both programmes dispense with giving intentionality and mental states any fundamental theoretical, descriptive or explanatory role. If, on the other hand, one embraces mentalism – as Freud, to his great credit, did – then experimentation will, in practice, never be enough (at least, not insofar as we can envisage developments in experimental methodology for the foreseeable future). For example, empathy is an aspect of the way in which we make inductive inferences about the mental states of others by FP-reasoning which has no counterpart in current experimental (or epidemiological) methodology. And it is possible that empathy could never be reproduced by such methods, even in the future. On the other hand, empathy is central to understanding in psychotherapy, as it is to effective interpersonal understanding in everyday life (see Bolton and Hill [1996] 129-139).

CHAPTER 3

OPPORTUNITIES FOR CASE-STUDY CAUSAL INFERENCE

- AN EXAMPLE

3.1 REASONS FOR FOCUSING ON GENERAL PSYCHOTHERAPEUTIC COUNSELLING (GPC)

The principal model of psychotherapy used in this thesis is General (or non-doctrinal) Psychotherapeutic Counselling (GPC), not Freudian Psychoanalysis (FPA). This is because there are some methodological and epistemological issues of importance which are better illustrated by GPC than by FPA. For example:

(i) The lines of continuity between FP and GPC (or between FP-reasoning and the reasoning involved in GPC) are undeniable, and much clearer than in the case of FP and FPA. Consequently, the lines of continuity between the way in which we (informally) test hypotheses about others (or ourselves) in everyday life and the way GPC counsellors do so for their clients is better illustrated. There are, moreover, some advantages of looking at this simpler case first, before (or in addition to) examining inferences in theoretically stronger psychodynamic systems. Because it is simpler (and uncluttered by special theory) certain basic elements of interpersonal psychological theorising and inference are easier to see – for example, the dependence on intentionalistic postulates and explanations. Also, the kinds of hypotheses that can be formulated in GPC sessions, their value, and the reliability with which they can be (informally) tested have, to the best of my knowledge, not been adequately examined in the philosophical literature (which has instead concentrated heavily on FPA).

(ii) FP and GPC involve the kinds of observational and inferential opportunities that all theoretically stronger (or 'doctrinal') psychodynamic psychological approaches begin with (even if they add further theory). There is therefore justification for regarding them as epistemologically more basic (even if theoretically less ambitious). GPC thus provides a better opportunity for examining

some of the epistemic resources common to all (and basic to all) psychodynamic psychological schools.

(iii) Some (though admittedly not all) problems concerned with the inadequacies (or failure) of canonical experimental methods in being able to test hypotheses of interest to psychotherapists are better illustrated by GPC than by theoretically stronger psychodynamic approaches. Once again, this is because of the nearness of GPC to FP. The special concern of GPC is with category II hypotheses (although, as with all psychotherapeutic approaches, GPC also utilises category I hypotheses). We need to examine the reasons why experimental methods do not function well in testing these categories, if this is indeed the case (see chapter 5).

In what follows in this chapter I shall provide, firstly, a primarily descriptive outline of some of the features of GPC and, secondly, a specimen GPC case-study of my own (the 'Tom' case).

3.2 AN OUTLINE OF GENERAL PSYCHOTHERAPEUTIC COUNSELLING

For Eliminativists such as Paul Churchland (Churchland [1989] chapter 1) or others sceptical about whether FP terms refer, the entire programme of seeking to enquire into human psychology on the basis of the framework of FP concepts and explanations would seem to be a non-starter. Such enquiry is predicated on the assumption that FP concepts (such as beliefs, desires, thoughts, emotions, actions) pick out psychological realities (behavioural ones in the case of actions) at some level of description. It is, moreover, predicated on the assumption that FP explanations perform genuine explanatory work. In this thesis I shall not attempt to respond to Eliminativist criticisms of FP knowledge. It will be assumed that, to a very large extent, FP terms do refer and that FP explanations are genuine. However, if Eliminativist criticisms are valid, then clearly any programme based on the assumptions of the ontological genuineness of FP will be weakened or undermined.

If FP terms do pick out realities and FP explanations are genuine, then FP stands as a rational and empirically-grounded psychology, albeit of limited scope. By using it and the inferential resources of FP-reasoning it ought to be possible to acquire psychological knowledge about another person (or oneself). Moreover, given the conceptual/theoretical and inferential resources of FP and its reasoning, there is no obvious reason why the knowledge which (let us assume) can be acquired by them should be limited to what can be obtained in ordinary social life. The latter imposes various restrictions of a pragmatic, cultural, aesthetic and intellectual nature on the extent to which and the accuracy with which the aforementioned resources can be applied. I shall now argue that GPC offers a more refined and thorough utilisation of these resources, and that this makes possible an increase in the depth, scope and accuracy of the FP understanding of others (or ourselves).

In GPC one works within the framework of concepts and explanations that are more or less standard in everyday social life (i.e. FP), but operates with them in a setting in which topics are aired which social propriety or the inhibitions of the

client normally prevent from being brought into the open. In addition, much greater care and attention is given to the quality of the observations and inferences which are made than would be the case in ordinary social life (see also section 2.2). The client is encouraged to speak openly and truthfully about topics of emotional or personal significance to him/her, but may also be encouraged to face up to issues which the therapist judges he/she is failing adequately to address. There is much greater attention to detail and thoroughness in following topics through than would normally be the case in ordinary social life. The therapist is trained to be a patient, attentive and sympathetic listener. He is also trained to be non-judgemental with regard to the core feelings, impulses or ideas which the client expresses and to treat them with 'scientific neutrality' (i.e. clinical detatchment)[23]. The client should feel confident of professional confidentiality and that he is disclosing his intimate thoughts, feelings or life-experiences to a neutral party trained to receive such details in a matter-of-fact but humane and sympathetic way.

The sense in which an extension of knowledge is possible under this format consists principally in the information which the client makes available about himself/herself by speaking openly and truthfully about matters which would otherwise not be disclosed or adequately objectified, or would not be made available for being recorded. Further, the client may be able to learn something additional about himself by putting his private thoughts and feelings into an objectifiable arena (i.e. by giving linguistic expression to them). By doing this they can be commented upon and put into juxtaposition with other thoughts and feelings of his (or with patterns of his behaviour) in ways he may not have previously been aware of. Thus, we may be able to learn more about others (or about ourselves if we are clients) than would be possible in ordinary social circumstances, even though we do not move outside the general framework of FP concepts and explanations.

It might be felt that the sense in which the above constitutes an "extension of knowledge" is so limited as not to be worthy of that title. To this the following may be said:

(i) It is not the supporters of the above stance who are claiming grand or dramatic extensions of knowledge by means of enquiries of this kind. What is being claimed is a *de facto* enlargement in the scope and depth of the understanding of another person (or of oneself) in broadly FP terms. If anyone should find that unimpressive, so be it. What counts is whether the claim is true.

(ii) It is possible that, occasionally, material of good reliability will be presented in a GPC session which is nevertheless quite unexpected from the standpoint of even an enlightened 'commonsense' understanding of human psychology. This will need to be accurately documented and will call for explanation and possibly the development of new theory. As mentioned earlier (see pages 24-25), there is no strict demarcation between the kinds of data that could crop up in a GPC session and those that could arise in a session of any one of the theoretically stronger doctrinal psychodynamic schools (such as FPA). It would, moreover, be utterly implausible to dismiss every occurrence of new or unexpected data as being the result of suggestion. GPC typically does not use free-association at all, or otherwise uses it only minimally. Also, in GPC there is an emphasis on encouraging the client to talk and to disclose material (and to do so at his own pace): it is from this that inferences are made. The therapist, moreover, is likely to consider various theoretical interpretations of the data. This contrasts with the portrayal of him as a person who is under the grip of a single theory which he presumes to be correct, and that he is forcing the interpretation of the data (or even what the client is disclosing) into its mould[24].

The basic justification for the claim that, to a substantial degree, inferences in GPC sessions have the potential for being reliable is the following:

(i) The kinds of inductive inferences (and tacit theory) used in GPC sessions will, in the main, be the same as (or very similar to) those which we use in FP-reasoning in everyday life. (Note, however,

that in GPC sessions therapists may augment the tacit theory of FP with specialist knowledge drawn from scientific psychology or psychiatry; and they may augment FP-reasoning with consciously applied normative methodological principles.)

(ii) The kinds of inductive inferences (and tacit theory) which we use in FP-reasoning in everyday life by and large have the potential for delivering reliable conclusions (i.e. those which are true or deserving of rational credence) over a certain range of hypotheses (i.e. category I). (Some arguments supporting this claim are provided in section 7.22)

(iii) Consequently, to a significant degree, conclusions drawn in GPC sessions by means of FP-reasoning have the potential for being true or deserving of rational credence.

I have used the word "potential", above, when speaking about the capacity for drawing reliable conclusions in both everyday FP and GPC contexts. This is because even though such a capacity exists (or is latent) it does not follow that it will always be effectively or successfully employed. Many factors (for example, individual differences in acquired skill or in latent ability at deploying FP-reasoning; inherent limitations in the severity of FP-reasoning 'tests'; insufficient or inaccurate data etc.) could affect the outcome. Moreover, the above argument addresses the question of the reliability of GPC inferences and conclusions only in a *global* way. That is, it does not address the problems attaching to the testing and validation of *individual* GPC claims (for the corresponding distinction with regard to FP see pages 61, 195 and 208-209).

I accept that the justification for GPC conclusions falls short of validation to ideal (or the highest scientific) standards. However, to suggest that, short of testing and validating category I and II hypotheses to the highest scientific standards, we can have no justification for believing them overstates the

acknowledged limitations and risks of carrying out the kinds of informal tests on them that we do in everyday life and in GPC sessions. Rational grounds for belief are, in general, better conceived as a gradation or continuum, rather than being an 'all or nothing' affair (see e.g. Strawson [1952] 237-238, 248). The failure of GPC and FP inferences to conform to the patterns and standards of mature scientific testing and validation at their best does not necessarily render all conclusions reached by them devoid of rational warrant. 'Global' arguments which counter the sceptic of FP-knowledge and which lend support to the view that we should expect the network of FP inferences and conclusions to be reliable to a substantial degree can be provided (see section 7.22). Moreover, specific arguments lending credibility to individual GPC conclusions (including causal ones) can sometimes be found (see section 3.34). While these may fall short of what would be required for the establishment of a scientific conclusion (in the orthodox sense) the degree of rational assent which they can require can be high.

3.3 A SPECIMEN GPC CASE-STUDY

3.31 INTRODUCTION

Given Grünbaum's strong stance regarding the need for controlled (i.e. experimental or epidemiological) studies to avoid the pitfalls of causal inference (e.g. Grünbaum [1984] 47, 48; [1986] 228) it might have been thought that he believed that such studies were *necessary* for that task. For example, in the passage quoted on page 22 of this thesis he does say that "controlled inquiries . . . are *inductively required* to lend genuine support to Freud's specific etiologies of the neuroses" (Grünbaum op. cit.); and elsewhere he says: "To guard against inductive fallacies of causal inference, methods of *controlled* inquiry are needed" (Grünbaum [1980] 368; emphasis in original). However, attributing to Grünbaum the strict necessity of experimental studies for testing and validating causal claims misrepresents his position, which is more moderate. Grünbaum certainly allows that some classes of FP causal inference (e.g. that insults can cause feelings of anger or humiliation, or that good tidings can bring about joy) can be known reliably without formal experimentation (see Grünbaum [1994](a) 54, quoted on page 29 of this thesis). However, in such cases Grünbaum maintains that some surrogate (my term) for controlled inference must be involved: for example, some tacit means by which Mill's methods are being implemented through the use of FP-reasoning (Grünbaum ibid.; see also Grünbaum [1993] 163-164). I think that the latter view is, in broad outline, entirely reasonable and just what we would expect if FP-reasoning does issue in reliable causal conclusions. (I do, however, take issue with the specific account of such casual inference that Grünbaum proposes in at least one case – see sections 7.3 and 7.4) Further, Grünbaum seems to allow that at least some causal hypotheses of GPC can be tested fairly effectively intraclinically. For example, he seems to accept an example (provided by Rosemarie Sand) in which an analyst inferred that a "teenage female patient's low self-esteem . . . was *caused* by her lifelong awareness of her father's vilifications of her" (Grünbaum [1993] 112; my

emphasis). Even further, it is possible that Grünbaum would not exclude the possibility that some distinctive hypotheses of FPA (including causal ones) could be tested with reasonable effectiveness in the clinical setting. Even though he does not make a straightforward statement to this effect, he does appear to cover himself against the claim that it would be impossible. He says, for example: "I do *not* maintain that any and all [psychoanalytical] clinical data are altogether irrelevant probatively" and seems to allow for "[the] existence of *some* circumstances under which we would be warranted in not renouncing [psychoanalytical] clinical evidence "altogether" " (Grünbaum [1984] 265; emphases in original). He also says: "I do not deny at all that *now and then* clinical results "are contrary to the expectations and belief of the clinician" " (ibid.; emphasis in original) and that his position "[does *not* amount to declaring] the automatic falsity of any and every analytic interpretation" (Grünbaum op. cit. 245). Also, referring to a passage by Edelson from the latter's [1988], Grünbaum says:

> "[Edelson] somewhat overstates my skepticism when he speaks of my "conclusion that any reliance by psychoanalysis on data obtained from the psychoanalytic situation to support its causal inferences according to ... [the] canons [of eliminative inductivism] is by the very nature of things doomed to failure and that psychoanalysis must turn instead entirely to epidemiologic and experimental research for such support." "
>
> (Grünbaum [1993] 254. The passage quoted by Grünbaum is from Edelson [1988] 275.)

Overall, therefore, Grünbaum seems to acknowledge that some GPC (i.e. category II) causal hypotheses can be tested intraclinically (i.e. without experimental controls) with reasonable or good levels of reliability; and even that there is a possibility (however slim) that this might apply to some FPA (i.e. category III) causal hypotheses. However, Grünbaum does not build on this or make a positive case for it. This should prompt us to ask the question: "Why not?". If there is *any* genuine scope for reliably testing causal hypotheses (of categories II or III) intraclinically, then this needs to be brought out into the open and advertised, and the extent of the facility needs to be examined and critically assessed.

In contrast to Grünbaum's apparent attitude, I think it is important to make a fully positive case for whatever capacity there is for making reliable or

reasonably reliable intraclinical causal relevance inferences. Not to do so might result in our failing adequately to appreciate the extent of the knowledge we are entitled to, even if the hypotheses concerned have not been tested by canonical methods of mature scientific practice, or to standards associated with the best scientific confirmations.

In what follows I shall, by means of an illustrative example, attempt to show how some causal inferences can be made in GPC sessions without the use of externally applied controls. The range of psychological themes which psychotherapists encounter is very wide. A single case-study, such as the one provided, can do no more than illustrate a few actual inferences and the kinds of inferential strategies involved. It is not being claimed that, even under optimal conditions, what issues from such intraclinical inferences are conclusions having the epistemic characteristics of science. But, in my view, this is not a fatal flaw so long as the level of reliability of those conclusions is high, and the hypotheses involved are important (theoretically or pragmatically). My main point is that, for many of the inferences, the level of reliability of the causal conclusions reached is greater than that which Grünbaum (and other critics) openly and prominently acknowledge. Moreover, in the most favourable circumstances, the level of rational credence attaching to those conclusions can be very high (beyond reasonable doubt).

Let us now turn to the case-study. The case of 'Tom' – as I shall refer to the person concerned – is based on the real case-history of someone known to me, and with whom I had the opportunity of discussing the material in considerable detail. I have altered some of the basic facts to conceal identity. However, none of these alterations, in my judgement, impair the lessons which I am attempting to draw from the case. These lessons are methodological or epistemological. I shall not be concerned with any psychological conclusions, *per se*, which can be drawn from the case, interesting though they may be. The lessons concern the feasibility and reliability of intraclinical causal inference. In the course of making the inferences very considerable use is made of 'background knowledge'. This exemplifies the central role which it is felt that background knowledge plays in clinical inference.

3.32 PRELIMINARY OUTLINE OF THE 'TOM' CASE

For a lengthy period of time, intermittently at first, and reaching a peak in his twentieth year Tom experienced a range of incapacitating psychological symptoms which included anxiety and depression. After consultations with both his general practitioner and hospital (and treatment by them with antidepressants) he was seen by a psychotherapeutic counsellor. It seems that some of the psychological symptoms that Tom experienced during this period may well have been of organic origin. He had been recurrently ill with bouts of bronchitis during part of this period, and it is not unreasonable to conjecture that some of his psychological symptoms were caused by this organic illness. However, Tom was adamant that not all of his depression or anxiety is likely to have been caused in this way. There were at least two broad and distinct emotional problem areas which were of continuing psychological burden to him. Both were already in place long before the onset of the bronchitis. It is therefore totally implausible that either of these 'emotional problem areas' could be explained as the effect (i.e. symptoms) of his bronchial illness.

The emotional problem area I shall be dealing with appears to relate to a complex web of social maladaptation beginning in Tom's early childhood and later involving very specific feelings of guilt towards his father (I shall not be dealing with the other problem area at all). According to Tom (and from his own subjective perspective of assessment) a large part of the therapeutic benefit that he gained from the psychotherapy sessions (which were of a GPC nature) lay in being given the opportunity to face and articulate his emotional problems and bring them out into the open. Neither his general practitioner nor the consultant at the hospital had begun to offer such a service. Moreover, he had not been able to air the problem area we shall be discussing in a fully adequate way with friends or family, and had not been able to discuss the other area with them at all.

I shall next sketch an outline of some of the main pertinent facts of the case, based on Tom's discussion of them with me. Later, (in section 3.34) I shall focus on some episodes in the case-history in greater detail. In each case a causal

hypothesis will be formulated, and we shall consider to what extent rational credence in that hypothesis is justified.

There was a key incident which occurred during Tom's first year at his primary school (when he was aged six). It had somehow come to be known amongst some of the children that Tom's father was an invalid (he had been severely paralysed following a car accident when Tom was very young and was wheelchair-bound). A few children, including a notorious school bully whom I shall call Dan, encircled Tom in the playground during one of the play intervals. Dan then came up to Tom in an intimidating way and asked: "What kind of father have you got?". Tom said that some of Dan's accomplices laughed in a mocking way and that he was taunted and questioned further. Tom said that in response to this interrogation he experienced a set of strong emotions which included fear and humiliation. He could not bring himself to admit to his peers (under these circumstances) that his father was an invalid and instead said (untruthfully) "He's better now". Tom emphasised that this event was extremely traumatic for him. He said that he came to fear the children involved. He also came to dread the possibility that anyone (apart from his family, their friends, and a few child friends who did not attend his school) would discover that his father was an invalid. Subsequently, throughout his school years he strove to conceal that fact from the pupils at his various schools, fearing that if it were found out he would be stigmatised or would be subjected to torments or even physical aggression. During the same period Tom's relationship with his father was an affectionate one. There thus arose a situation in which, outside the circle of his family and a few friends, Tom concealed the fact of his father's invalidity and took active steps to prevent that fact from becoming known. For example, Tom never invited children, including friends, from his school to his home and actively discouraged them from visiting. On the other hand, Tom's bond of affection towards his father was strong, and when he actively prevented his peers from meeting or finding out about him he experienced feelings of guilt for, as he put it, having been unable to accept his father "as he was" and "for himself" in the presence of others.

3.33 THE RELIABILITY OF THE DATA

Critics of psychotherapeutic clinical inference standardly raise the unreliability of the data as their first line of objection against the claim that reliable conclusions can be drawn from them. Consequently, we need to address this potential objection with regard to the present case. The two main potential sources of the vitiation of the clinical data which Grünbaum raises for FPA are: (i) the suggestive influence of the analyst (see e.g. Grünbaum [1984] 130-131, 212-215, 244-245, 264); and (ii) the unreliability of the patient's memory (see e.g. Grünbaum [1984] 242-244; [1993] 249-250). Let us consider these in turn for the case of 'Tom'.

(i) In Tom's case the risk of any significant distorting suggestive influence must be regarded as very low. According to Tom, his psychotherapeutic counsellor aimed primarily at encouraging him to talk about whatever might be burdening him emotionally. No theoretically strong interpretation was put on what Tom had to say – it could be understood in broadly FP terms. Further, no use was made of free-association. There was conversational 'exploration' by the therapist of what Tom had to say but, according to him, no conclusions were forced upon him. He was encouraged to face the emotional problems that were troubling him (rather than to flee them), and to be open and truthful in talking about them. However, he was left in charge of what he chose to say or not say.

Tom said that probably the greatest difficulty he had was simply in being able to 'open up' and talk about themes that were linked to the anguish he felt. He was, however, already aware of some of the core aspects of these themes, and had been for a long time. At the same time, he said that he could not envisage how talking about or exploring these themes could make any difference to the general - and incapacitating - misery that he felt. The therapist, moreover, made no promises about what therapeutic gains might be achievable[25].

Overall, with regard to those details of the Tom case that we shall be discussing, I think there is no good reason to believe that *any* of those details are spurious artefacts of a suggestive influence due to Tom's therapist.

(ii) The sample causal conclusions that we shall attempt to draw from Tom's case (in section 3.34) do depend crucially on the veridicality of Tom's memory, since his memory supplies the principal source of the data that we shall be using. Are we justified in relying upon his memory?

Certainly, there is at present no reliable scientific method of checking whether a long-term memory (or any memory) is correct. Such a method would have to provide us with an independent and objective (as well as readily applicable and effective) test of whether what a person subjectively claims to have occurred on the basis of his/her memory actually did occur. We shall, for present purposes, ignore fortuitous external circumstances which, by chance, facilitate the confirmation or refutation of a claim made on behalf of a memory – that does not constitute a method. A standard view is to disallow long-term memories from counting as bases for making claims deserving of scientific credibility. This amounts to a position of epistemic caution. It is tantamount to the principle that we shall refrain from giving credence to that which cannot be independently tested and objectively demonstrated. Grünbaum appears to adopt a cautionary attitude of this general kind when he says:

> "In the court of science, credence should not be given to the unsubstantiated beliefs of analysts that they have an intuitive ability to discriminate pseudo memories from authentic ones – the less so, since their intuitions are bound to be steered and regimented by the theoretical exigencies of their craft in the context of their therapeutic objectives."
> (Grünbaum [1993] 250)

However, whereas the above principle is virtuous in that it functions to exclude what is false, it is easy to see that it is epistemologically defective. This is because the principle, if applied, would not only exclude the (false) information supplied by false long-term pseudo-memories, but also the (true) information supplied by true long-term memories, if there are any. That there are some memories of the latter kind cannot be ruled out. It is not merely logically possible but contingently possible (and perhaps cognitively actual) that some long-term memories do veridically attest to events in the distant past. The aforementioned principle does not, therefore, discriminate between false long-term (pseudo-) memories, deserving

of exclusion, and true ones which cannot be independently tested or objectively demonstrated, but which are deserving of acceptance because they are true (see also section 2.4). Moreover, since the information supplied by veridical long-term memory could be of very considerable psychological importance (both theoretically and practically) it is essential that any such information, if it exists, is not lost but retrieved.

Just how reliable is Tom's memory regarding the events described in the outline above (and in the more detailed analysis, below)?

Tom himself insisted that the details as he presented them were in all essentials correct. If asked a question (or if a doubt was raised) he gave every impression of trying to make the truth known, freely admitting lack of knowledge or uncertainty where appropriate. I detected no inconsistencies in his account. His responses to questions were natural, fluent and direct, conveying a sense of someone who was subjectively in touch with memories of past realities. There was certainly nothing to suggest that Tom was generally deluded or that he was trying to concoct a story. On matters which were of emotional significance he spoke with a sense of earnestness. One certainly received the impression that he was speaking about events that had deeply affected him. Moreover, if one doubted his sincerity or the accuracy of his account it would be necessary to explain why he insisted that events had taken place just as he had described them, and returned again and again to provide the same details. These details were not merely unvarying, but were consistent with one another. This was so even though the statements in which they were given appeared in different contexts on different occasions (including on separate days over a lengthy period) as different aspects of the case were discussed from different angles.

It is certainly possible that someone with a false memory could believe it to be veridical and could insist that it was so. It is also possible that on some points of detail Tom could be mistaken (though it should be added that I had no specific reason or evidence to suspect this). However, even if one sceptically assumes that some of the details in Tom's account are false I do not think there is good reason to doubt it on any of the main points. Thus, although the data Tom provided was not scientifically authenticated it has, in my judgement, the hallmarks

of being in all essential respects accurate and reliable. I could not find any specific reasons for seriously doubting any of the principal claims made.

Of course, the subjective indicators mentioned above which would tend to lend credence to the authenticity of Tom's account cannot serve as a substitute for its scientific authentication. It has also been acknowledged that there currently exists no (scientific) method which could provide an independent check on the data's authenticity. It could therefore be insisted that since the data is not (and possibly cannot be) scientifically authenticated it should not be used. But this, I believe, would be unduly to restrict what data is deemed permissible. Outside science (and possibly even sometimes within it) we have to rely on data that is not authenticated to scientific standards. This does not imply that the data is untrustworthy. Given the factors mentioned above which lend credibility to the reliability of the data, it would seem unreasonable not to admit it.

3.34 INFERRING CAUSAL CONCLUSIONS – SOME EXAMPLES

In this section we shall examine some examples of causal hypotheses that can be formulated about the material provided by Tom's account. They will be labelled T1 – T4. In each case our principal concern will be to evaluate whether inferring a causal connection (i.e. a positive causal relevance relation) is epistemically justified.

Hypothesis T1:

> 'The intimidating and interrogating behaviour of Dan (and his accomplices) during the playground incident was (positively) causally relevant for Tom's fear and general distress during that incident'.

How reliably can the truth or falsehood of this causal hypothesis be established (given, as we have judged it to be, that Tom's report is descriptively accurate)? In my view, by relying on background knowledge it can be ascertained with a very high degree of reliability, and affirmatively. That background knowledge includes: our 'commonsense' (but, it is to be noted, not scientifically tested) understanding of the effects of aggressive, derisory behaviour[26]; the fact that Tom is a sensitive person (and on all accounts was a sensitive child); that he was only six years old (his sixth birthday was within a few weeks of the event); and that (as he informed me) he joined the school only a few weeks prior to the playground incident and found his surroundings unfamiliar and isolating. We would expect (as an instance of an empirical generalisation for which there was evidential support from everyday life) that, *ceteris paribus*, a sensitive child would be traumatised by an incident such as the one described. Tom's own testimony was that he found it very distressing: he was emphatic in regarding it as one of the most traumatic events in his life. But might not the inference of a causal connection between the behaviour of Dan (and the others) and Tom's distress be a *post hoc ergo propter hoc* fallacy? In my judgement this is so implausible as to be, in effect, discountable. How likely is it

that Tom would have experienced distress at precisely the time and place that he did (having been surrounded and aggressively 'interrogated' by Dan and his cohorts) and yet it *not* be the case that the action of Dan and the others were positively causally relevant for that distress? If such an event did take place it would be a remarkable phenomenon indeed. Not only would we need to explain why Tom felt distress of the kind that he did at precisely the time and place that he did (and, indeed, there were thematic links connecting Dan's utterances and part of what Tom was distressed about – see Hypothesis T2, below); but we would also need to explain – contra all reasonable expectation – why Dan's focused aggression was having no causal influence on Tom's psychology despite its spatio-temporal proximity. The likelihood of there not being a causal influence between the behaviour of Dan and the others and Tom's distress during the key incident is virtually zero.

After the key incident Tom was the recipient of further intimidating behaviour and remarks from Dan (but he was not the victim of any actual physical violence). He developed what it might be appropriate to describe as a phobia of encountering Dan (or any of Dan's accomplices) either in school or outside it. Tom developed a specific anxiety about going into the playground during the play-intervals when he would be compelled to share the same play area as Dan (Dan was a year or two older than Tom and was in a separate class). We can ask causal relevance questions about these anxieties and behavioural tendencies. For example, how likely is it that Dan's aggressive behaviour towards Tom was (positively) causally relevant for Tom's phobic behaviour? Once again, I think that on background knowledge we can infer a positive influence with a very high level of confidence. This can be done, once again, by way of asking ourselves a hypothetical contrast-class plausibility question: given background knowledge, how likely is it that those (specific) fears would have arisen in the specific circumstances that they did, and it *not* be the case that there was a causal influence?

So far we have concerned ourselves only with some fairly general features of Tom's emotions and behaviour during and subsequent to the playground incident. I would, however, like to consider details pertaining to cognitive aspects of Tom's emotions. Dan and the others were not only generally aggressive and

intimidating, but quizzed Tom on a specific issue which, by making it the theme of their 'interrogation' drew attention to it and made Tom conscious of it in a heightened way. That theme was, of course, Tom's father's invalidity. The attitude which Dan expressed towards that fact was one of aggressive derision. How might the forceful expression of this attitude have affected Tom's awareness of that fact and his emotional reaction to it?

According to Tom's own account, during the key incident Dan and the others made him tacitly aware (through their behaviour and remarks) that they regarded severe disability with disdain, and that because he was associated with such disability that they regarded him as a suitable target for mockery and even hostility. It was, indeed, awareness of these attitudes which, Tom said, led him effectively to lie about his father's condition by saying "He's better now".

With this background, let us formulate an hypothesis pertaining to the possible causal influence of the specific content of Dan's remarks on Tom's mentation (both emotional and cognitive) during the key incident:

Hypothesis T2

> 'The specific issue (or theme) in respect of which Dan interrogated Tom during the key playground incident was positively causally relevant for Tom's feeling humiliated and anxious *about that same theme,* (i.e. his father's invalidity) during the incident.'

How likely is it that T2 is true? How likely is it that Dan asked Tom questions pertaining to his father's invalidity (which he did), that Tom was aware of the questions and reacted to them both cognitively and emotionally (which he did), and yet it *not* be the case that there was a causal influence from the content of Dan's utterance to the content of Tom's subsequent thoughts, as well as to the cognitive content of Tom's emotions (i.e. what it was that he felt humiliation or anxiety about)? On a materialistic theory of the mind (which is what is being assumed here) the relaying of information (encapsulating semantic content) between two agents has

to take place on a physical substratum and involves a causal process. On the assumption of such a theory, it is therefore implausible that Dan will have uttered the specific question(s) that he did, that Tom will have responded specifically (i.e. in a manner which matches that content) and yet there have been no causal connection. This is not a case of begging the (specific) causal question because the assumption of a general causal nexus is an entirely non-specific background metatheoretical (or metaphysical) one. The reasoning behind our specific claim is probabilistic and based on plausibility. Given that there was (as a matter of fact) a matching of content as described, it is extremely unlikely that the content of Tom's response (cognitive-emotional) will have been as it was and yet not have been causally influenced by the content of Dan's questioning[27]. We therefore have good reasons for believing Hypothesis T2 to be true.

T2 differs from T1 in that T2 includes an implicit reference to meaning: i.e. the meaning connected with the specifiable theme in respect of which Dan derisively interrogated Tom. There is evidence to suggest that the significance of that theme was not lost on Tom but continued to influence his mentation and behaviour for a very long time after the 'key' incident. There is little doubt that for most of Tom's later childhood and adolescence he was motivated to act in a way that was defensive about, specifically, it becoming known to others that his father was an invalid. There is, of course, a theoretical question concerning the exact role (if any) that the key incident played in giving rise to that propensity. It is possible that other events which Tom had forgotten about, or even general and pervasive negative cultural attitudes towards disability also contributed towards its development. Furthermore, it is very likely that factors related to Tom's own personality played a crucial role. For example, a child less timid or sensitive might have reacted in an altogether different way (e.g. by being aggressive and derisive towards anyone who was so towards him; or by not attempting to conceal the target of derision but proclaiming it defiantly). It has to be admitted that the inferential opportunities of case-studies almost certainly do not permit an accurate ascertainment of the respective contributions of these (or other) factors. However, all we need to consider here is whether, given that Tom's personality as a child was sensitive and somewhat

timid (this is what the evidence suggests), did the key incident make a significant (positive) difference with regard to the development of the defensive propensity? Given background knowledge, I think that we can infer with a reasonable to good level of confidence that it did. We would expect that the key incident taught Tom that having a severely invalid father made him vulnerable to taunts and aggression from some individuals, and that if that fact was not known then the potential source of such hostility was removed. This in turn was a powerful motivator (conscious or otherwise) for concealing the fact. It certainly seems implausible to claim that the key incident made no difference with regard to the development of the defensive propensity, or that its effect was ameliorative.

Before proceeding to Hypothesis T3 it will be necessary to provide some further case-history information. Shortly over one year after the key incident in the playground Tom was invited to the birthday party of one of the children at his school with whom he had become friendly (Bill). The event can be dated accurately because it was for Bill's seventh birthday, about six weeks after Tom's. Towards the end of the party, when parents were beginning to collect their children, Bill's mother asked Tom when his parents would be coming to collect him.

Tom said that he became very anxious on being asked this question because of an awareness that whatever he said it might inadvertently 'give away' the fact of his father's invalidity, and he dreaded this becoming known[28]. (N.B. Bill's mother did not know about Tom's father's invalidity.) However he replied, he might be asked further questions: such as which of his parents would be collecting him, if not both, and why his father would not be coming, and so on. This, at least, was the kind of avenue down which Tom said he sensed – and feared – the conversation might lead[29]. Tom said that the outcome of this was that he did not answer Bill's mother directly and became awkward and tongue-tied, to her bemusement. Tom was emphatic that he had a distinct and reliable memory of the event depicted above, including his own motive for stalling in replying to Bill's mother's question (i.e. that whatever he said might reveal, or might lead to further questions which would reveal, his father's invalidity).

Should we be allowed to rely on Tom's claim of having a veridical memory of his motive for not replying directly, given that the event referred to would have been approximately 15 years in Tom's past, when he was aged seven? In his [1984] Grünbaum criticises and, in effect, rejects the reliability of long-term memories (in his view, merely allegedly) recovered in FPA case-studies (op. cit. 242-244, 257). Especially in view of the fact that that work has had considerable influence, it needs to be emphasised that Grünbaum in no sense establishes the general lack of veridicality of long-term memories, even for events in infancy remembered in adulthood. This is not to deny that there are legitimate grounds for scepticism when a pernicious suggestive or distorting influence is at work. However, it is possible that there are many occasions when long-term memory operates veridically in the lives of ordinary individuals (including, of course, those *not* seeking psychotherapy). Grünbaum provides no conclusive arguments or evidence for ruling out this possibility on empirical or epistemic grounds, and it remains entirely an open matter whether veridical long-term memories are sometimes a cognitive fact. In keeping with the points raised in the earlier discussion (section 3.33(ii)) I maintain that there are good grounds for believing that Tom's account is in all essential respects reliable and accurate and shall proceed on this basis.

T3 can now be formulated:

Hypothesis T3

> 'Tom's motive of not wanting to impart information that would lead to its becoming known that his father was an invalid was (positively) causally relevant for his not replying directly to Bill's mother's question.'

It is assumed that Tom (or any average seven-year-old) has the capability of answering in a direct and unhesitating manner any question put to him/her which falls within the range of his/her comprehension and competence to do so. Bill's mother's question falls into this category. Why then did Tom become

anxious, as well as tongue-tied and embarrassed when asked it, and why did he not answer directly? The answer is given by Tom himself, through a memory of his own motive at the time. His desire not to reveal the fact of his father's invalidity (inadvertently even if not directly) was the motive behind his prevarication. However, even if we accept this explanation as reliable and accurate a further problem lurks.

Let us keep in mind that our central concern is with the problem of establishing causal relations. Even if the above FP motivational explanation for Tom's prevarication is regarded as correct it still leaves a causal connection less than unequivocally established. The problem is not peculiar to our example, but is a general one, pertaining to an entire class of FP motivational explanations in which motives are provided as explanations for actions. In the 1950s, and stemming from a tradition which includes nineteenth century hermeneutic thought as well as some views of Ludwig Wittgenstein and R.G. Collingwood, a contrast was standardly drawn between motivational explanations involving reasons (for action) and, on the other hand, causal explanations (Moya [1990] 105). In that tradition explanations involving reasons or motives were viewed as non-causal (Moya ibid.). However, largely due to the work of Donald Davidson (especially Davidson [1963]) a turn-around on this issue occurred. Davidson argued that FP motivational explanations for actions are typically causal[30]. There is now a fairly large consensus that explanations of actions in terms of beliefs, desires or other mental states are legitimate causal explanations (Guttenplan [1994] 80). It is worth noting that even Grünbaum unambiguously accepts this[31]. For example he says:

> "Clearly, if an agent is actually moved to do A by having a certain reason or motive M – so that his having M explains his action A – then this very presence of M made a difference to his having done A. But, in that case, the agent's having M qualified as being *causally relevant* to what he did, *regardless of whether M is conscious or repressed.*"
> (Grünbaum [1984] 72; emphases in original.)

However, even if we accept that FP motivational explanations can be causal, and even if we accept (as I provisionally do) that the causal relata in such explanations are (i) beliefs, desires or other mental states (as potential causes) and

(ii) actions (as potential effects), there still remains a problem with regard to the hypothesised causal relation(s) being established to sufficiently high standards. Philosophical arguments, however persuasive, for regarding desires and beliefs as capable of being causally relevant for actions, coupled with examples of 'good' (i.e. compelling) real-life motivational explanations, certainly do not together supply grounds for the scientific validation of the putative causal relation(s). There thus exists a problem as to precisely what causal relation(s) we should be allowed to infer from FP motivational explanations (even when the explanation is 'good'), and what degree of rational credence should be allowed to attach to any causal relation that is inferred. One critic, Edward Erwin, has implied that the putative causal relation(s) in such cases need to be "established" (Erwin [1993] 446), and complains that the causal claims implicit in FP motivational explanations typically beg the question (op. cit. 446-447) (I discuss his comment in section 7.21). In the present case we have to concede that, even if we take Tom's explanation for his prevarication at the party to be reliable (and correct, as far as FP motivational explanations permit), our claim to be able to infer from this a causal relation must be fairly modest. It will be of a kind comparable to that to which we would be entitled in good (compelling) cases of intentional explanations in everyday life. As has already been stressed, this is weaker than claiming that the causal hypothesis has been "established"; or that it has been independently and objectively tested; or that it has been validated to scientific standards.

Prior to formulating T4, let us recapitulate that throughout much of his later childhood and most of his adolescence Tom was averse to being seen in public with his father and sometimes acted to prevent it. During this period he also dissuaded children (including friends) from school from visiting his home for the specific reason of preventing them from meeting (and hence seeing) his father. At the same time Tom's relationship with his father was an affectionate one. His desire not to be known as having an invalid father (or his active steps taken to prevent that fact becoming known) were often accompanied or followed by feelings of guilt.

T4 can now be stated:

Hypothesis T4:

'Tom's desire and repeated efforts, during his school years, not to be seen in the company of his father or to be associated with him was (positively) causally relevant for his feeling of guilt towards his father for having (in his own words) "betrayed" or "disowned" him.'

In this case there is a complication which also allows us to illustrate some of the limitations of GPC case-study inferences.

Tom experienced guilt (for desiring not to be seen or associated with his father) under two circumstances which need to be distinguished: (i) prior to the onset of his neurotic disorder; and (ii) after the onset of his neurotic disorder (and during its course). The worst phase of his neurotic disorder (which was after he had left school) was also after he had *stopped* trying to prevent himself being seen or associated with his father. During the very end of Tom's time at school – and especially after he had left school – he made an effort to set aright the schism which had developed between the public appearance of his not having a disabled father and the reality. He no longer took any measures to prevent himself from being seen or associated with his father, and even to some extent encouraged the gaining of knowledge by his peers of the truth. However, it was during this second phase that the feelings of guilt were most severe. In one sense this is not surprising. Exaggerated, 'out of control' emotions, arising 'out of context' of the immediate external circumstances which would make them rationally understandable (or justified) are a common feature of various psychological disorders. On the other hand, this does pose a problem with regard to the credibility of the hypothesis that Tom's desire (and active efforts) to conceal the crucial fact were the primary cause of the guilt he experienced as part of his *neurotic* phase. Simply put, when he was most keen (and active) in concealing the crucial fact he did not suffer a neurotic illness; when his neurotic disorder was at its height he was not (and had not been for some time) attempting to conceal the crucial fact. The hypothesis that his desire and

efforts to conceal his father's invalidity were primarily responsible for, specifically, the *neurotic guilt* therefore seems untenable. Nevertheless it is, I believe, still very probable that those factors were (positively) causally relevant for his *ordinary guilt*, experienced prior to the onset of his neurotic disorder and following thoughts or acts of disownment. We need therefore to distinguish 'ordinary' and 'neurotic' species of guilt. In Tom's case both species involved guilt at having rejected or disowned his father. Whereas an explanation of the ordinary guilt in FP terms is, I believe, highly credible, an FP explanation of the neurotic guilt is, by itself, very unlikely to be tenable. For purposes of clarity let us then distinguish two hypotheses which are not separated in T4:

Hypothesis T4A

> 'Tom's desire and repeated efforts (while he was at school) to prevent himself from being seen or associated with his father was (positively) causally relevant for the feelings of 'ordinary' guilt which he experienced (during or soon after those occasions) for having "betrayed" or "disowned" him.'

Hypothesis T4B

> 'Tom's desire and repeated efforts (while he was at school) to prevent himself from being seen or associated with his father was (positively) causally relevant for the neurotic guilt which he experienced for having "betrayed" or "disowned" him (this was part of his neurotic illness).'

An FP explanation of the ordinary guilt is very credible. There is a widely accepted empirical generalisation of FP to the following effect: *ceteris paribus,* a person who desires not to be associated with (or actively dissociates himself/herself from) someone with whom he/she has a strong and ongoing bond of

affection will experience guilt about, specifically, the betrayal of that person. There is supporting evidence from everyday life for this empirical generalisation. For example, we are inclined to experience guilt if we are disloyal to or dissociate ourselves from those to whom we are affectionately attached (so long as there is no special circumstance which justifies or excuses such a dissociation on a particular occasion). The guilt is, moreover, specific and intentionally directed: it is guilt at having betrayed or 'turned one's back on' someone who has loved or been loyal to us. In such cases there has, of course, been no independent (or scientific) test of the causal relevance relation (between disowning and guilt), yet it is not unreasonable to suppose that such a (positive) relationship holds. The 'informal' testing of everyday life seems to confirm the view that guilt of this type is experienced when disloyalty or disowning is present, not experienced when we are accepting of the loved person (even if it is at a cost to ourselves), and even that guilt can evaporate if we change our attitude and behaviour from those of disloyalty or dissociation to affiliation and acceptance (analogous to tweaking the independent variable in a controlled experiment)[32].

There were certainly occasions, recounted by Tom, on which he instantiated the antecedent of the aforementioned empirical generalisation (in connection with the specific theme of not wanting to be known as having a disabled father). There were also occasions during the same period (i.e. while he was at school and before the onset of his neurosis) when he experienced intentionally-specific guilt of the kind expressed in its consequent. In these cases, the FP explanation in terms of the widely accepted empirical generalisation provides not merely a plausible explanation of Tom's guilt but a fairly compelling one. On background knowledge, how likely is it that Tom would have desired not to be seen or associated with his father (and sometimes even acted to facilitate this) and yet not feel guilty about it? We have moreover, evidence from Tom's memory informing us that he did experience guilt during that period for just those reasons (in the intentional sense). Overall, I think we are justified in concluding with a good level of confidence that T4A is true. However, as with T3, caution and modesty are called for in drawing causal conclusions from FP explanations.

Whereas T4A dealt with 'ordinary' guilt, T4B deals with the guilt which Tom experienced during his neurotic illness. This was when Tom was no longer attempting to conceal the fact of his father's invalidity or, at least, had made a considerable effort to overcome that tendency. In this case it is very doubtful that the FP system of explanation can supply an adequate account. In this case the neurotic species of guilt was present even when the antecedent in the aforementioned FP empirical generalisation was not instantiated (or was severely weakened). There seems to be nothing in FP explanatory strategies that could show what made the crucial difference between the occurrence of the guilt in the ordinary and the neurotic cases. For the latter what is needed is new theory – theory which almost certainly goes well beyond that which is tacit to FP. At the same time, it would seem to be a remarkable coincidence if there were no causally continuous process between Tom's ordinary guilt and his neurotic guilt. The reason for the latter claim is that the intentional content of Tom's guilt – i.e. what he felt guilty about – was the same in both cases. This requires explanation. If the development and persistence of his neurotic disorder were causally autonomous one would not expect this. One possibility is that whatever are the specific causal factors responsible for the emergence of the neurotic state *per se*, such a state (of neurosis) will be 'parasitic' on emotional vulnerabilities which were set up in 'ordinary' (i.e. non-neurotic) conditions prior to the neurosis.

3.4 CONCLUDING REMARKS

It will be obvious that T1-T4A were not tested by being the subjects of experimental set-ups specifically designed to test them. This contrasts with the standard situation in empirical science in which, after an hypothesis has been proposed, an appropriate experimental strategy is devised and implemented to test it. Does this difference render the informal test-situation in the case-study epistemically 'null and void' (or make it methodologically inferior to an experimental test of T1-T4A)? I think that any evaluation of the epistemic benefits of controlled experimental tests, considered in the abstract, has to be weighed against whether such tests of T1-T4A are realistic in practice. In the abstract, one can consider the advantages of controlled experiments over *post hoc* or retrospective assessments of causal relevance of the kind that were in fact made for T1-T4A (and note that nothing more was ever claimed for the "testing" of T1-T4A than that they were being evaluated informally in this way). Considered abstractly, carefully controlled prospective experiments do have certain advantages over retrospective inferences. These are well-recognised in the psychological and medical literature (see e.g. Miller [1984] 4), and Grünbaum makes a point of emphasising them (Grünbaum [1983]; also see Grünbaum [1984] chapter 8; [1993] chapter 8)[33]. The principal advantage of prospective studies over retrospective ones is that they permit (in principle) the experimental control of the effects of extraneous variables. However, in the final analysis testing (as with the acquisition of knowledge in general) is constrained by pragmatic factors. Testing strategies (and the logic of testing which they implement) are of no use if they cannot be implemented *in practice*. In the present case, if no well-controlled or prospective experimental set-ups can be deployed in practice to test T1–T4A, then the canvassing of the epistemic advantages of experimentation (or prospective studies) amounts to nothing more than an empty – i.e. practically *un*realisable – gesture. The latter is, I believe, the situation that actually transpires in the case of T1-T4A. There is, it seems, *no* practically realisable experimental set-up capable of directly testing any of T1-T4A by means of carefully controlled or prospective strategies. The retrospective assessments of possible causal relevance in

each of T1–T4A, while falling short of ideal epistemic standards for tests (e.g. with regard to severity), nevertheless fare better than any practically realisable prospective experimental test. Those assessments (unlike controlled experimental inferences) can at least be made – and on carefully argued rational grounds.

A second possible criticism concerns the way in which the causal hypotheses in the case-study were constructed. It may be objected that they were simply 'phrased around' the information (i.e. details of life-history, experiences etc.) supplied by Tom, rather than being hypotheses in any genuinely inventive sense (as one might expect in *bona fide* empirical science). It is true that the causal hypotheses T1-T4A were tailored to the information supplied. However, it is not clear why this of itself should in any sense impugn them from the point of view of epistemic legitimacy (or even potential theoretical or practical value). As causal hypotheses, T1-T4A are as legitimate as any others[34] (it is, of course, a separate matter as to whether they - or others like them - can be *tested* as effectively as any others). Moreover, it is not evident that the circumstances of their construction are altogether unlike the way in which some hypotheses in *bona fide* science are constructed (at least, in some of the more applied – rather than purely theoretical – fields). A chemist who has acquainted himself with information from his laboratory manipulations and observations might, in conjunction with background knowledge (including various physico-chemical theories), hypothesise that factor (or variable) A was causally relevant for B. This has some similarities to the way in which a GPC counsellor might construct a causal hypothesis on the basis of information supplied by his client and background knowledge (although the kinds of data and background theory in the two cases are obviously very different).

If any genuine causal information is capable of being extracted from a case-study it will typically be without the implementation of any specially applied experimental controls (since standard case-study inferences do not employ them). How, without controlled experimentation can a credible case be made for the reliable acquisition of causal knowledge from a case-study? The answer is: by means of

arguments. It is the strength (or otherwise) of the arguments provided, on the basis of the details of the case, background knowledge, probability (or credibility) considerations, and the ruling out of alternative hypotheses that lend whatever degree of credence to the conclusions reached that they possess. For examples T1-T4A that degree of credence falls short of what would be ideal but, in my judgement, it is still high enough to warrant an overall acceptance of the hypotheses.

CHAPTER 4

SOME PROBLEMS WITH EXPERIMENTATION,

AND THE FAILURE

ADEQUATELY TO ACKNOWLEDGE THEM

4.1 INTRODUCTION

In this chapter I shall focus on two main themes:

Firstly, in 4.3 I shall examine some problems associated with experimental (or epidemiological) methods. My coverage is not intended to be systematic or complete. The focus will be primarily on certain group-comparison designs because these are favoured by Grünbaum and Eysenck and are also methodologically *de rigueur* in clinical and psychological research (Kazdin [1980] 103 ; Klerman [1986] 245).

Secondly, in 4.4 I shall argue that Grünbaum and, to a lesser extent, Eysenck do not draw adequate attention to the limitations and inherent problems of experimental and epidemiological methods in their writings on psychoanalysis. This is even though those methods are suggested by them as the means of improving upon or even overcoming the problems of psychoanalytic case-study inference.

In addition, at the end of the chapter (in 4.5) I discuss the plausibility of the suggestion that an hypothesis of the kind which Freud puts forward in his 'Rat Man' case - and which Grünbaum argues needs to be tested by prospective group-comparison methods (Grünbaum [1984] 257-261) - can, in fact, be adequately tested by such methods.

The above topics are prefaced in section 4.2 by an expository account of research designs which will provide a background for what follows.

4.2 RESEARCH DESIGNS

Standard classification of experimental designs in clinical and psychological research distinguishes Between-Group (also called 'group-comparison' or 'extensive') designs and Within-Subject designs (Kazdin [1980] 102-106):

(i) Between-Group Designs

In these different groups of subjects receive different treatments (i.e. experimental interventions or placebo) and the phenomenon of interest is studied by how it varies between the groups (Kazdin op. cit. 103-104; Chassan [1979] 147). Comparison between groups is made in terms of comparing the averages of the respective group scores (Chassan op. cit. 208).

(ii) Within-Subject Designs

In these each of the subjects receive all of the treatments, though these are administered in different temporal orders to the different individuals or groups of individuals. The subject serves as his/her own control (Kazdin op. cit. 104, 120, 160). Within-Subject designs traditionally use several subjects (e.g. a group, or several groups, each individual of which is monitored at successive temporal stages). However, a special case of Within-Subject designs is what Kazdin (op. cit. 105-106) calls the 'intrasubject-replication' design (also called the 'single-subject' or 'intensive' design). In the latter only one subject may be involved, the aim being to focus on the attributes or responses of that single individual. The most common single-subject design is the ABAB design in which two treatments (A and B) are alternated sequentially and the effects monitored (see Kazdin op. cit. 106, 172-176; Chassan op. cit. 226-227).

Whereas all experimental designs face problems (theoretical or practical), in this chapter I shall confine the discussion to some of those facing group-comparison designs only. This is for two reasons:

Firstly, as Kazdin ([1980] 103) has said:

> "The between-group design usually is considered as the standard against which the other design strategies are compared."

Consequently, in considering group-comparison designs and, especially, the Randomized Clinical Trial or RCT (see below) we are considering what most clinical researchers regard as the most effective means of detecting a causal influence, if there is one.

Secondly, Grünbaum expresses an explicit preference for group-comparison designs for psychoanalytical research (Grünbaum [1993] 243). Consequently, in addressing them we are confronting what Grünbaum presumably regards as the most effective means of potentially overcoming the liabilities of psychoanalytical case-study inference. A preference for group-comparison studies (RCT's) is also expressed by Eysenck (e.g. [1985] 46-49).

In the remainder of this subsection I shall provide an expository account of the three group-comparison formats most commonly used in clinical research. Only the third of these – i.e. RCT's – are truly experimental (and there are many possible variations on its design). Case-control studies and cohort studies are quasi-experimental. That is, whereas test conditions permit some discrimination between the presence/absence of a factor and its possible effect, strong controls in the manner of a true experiment are lacking (Campbell and Stanley [1963]).

(i) Case-control Studies

Within epidemiology one of the main study formats involving the comparison of groups is the case-control study. Backett and Robinson ([1992] 203) say of it: "The case-control study is probably the most widely used method to test an aetiological hypothesis". In essence it involves collecting a group of patients with a particular illness, looking in detail at factors from the patients' past histories which conceivably might have a bearing on the their condition (this obviously involves making use of a lot of background knowledge and judgement), and then seeing what the frequencies of occurrence of those same factors are in a control group who do not have the illness. Since the aim is to identify factors causally relevant for the

condition in question in terms of variables which have already had an effect, the case-control study is regarded as retrospective (Backett and Robinson ibid.; Alderson [1983] 128).

Although case-control studies are commonly thought of as hypothesis-testing, they are frequently involved in the generation of hypotheses, since ideas about aetiology might appear when comparing the data from the two groups (i.e. the 'case' and 'control' groups). This, I think, nicely illustrates that no matter how strongly some philosophers of science have insisted upon a fundamental logical and epistemological distinction between, respectively, the Contexts of Discovery and of Justification (see e.g. Reichenbach [1951] 231), the actual practice of learning about a target domain may not involve a clear partitioning of one 'context' from the other in procedural and epistemological terms. In employing both case and control groups the case-control study utilises aspects of the rationale of the Context of Justification – that is, it utilises the logic of contrastive evidential relevance. However, by searching through the life-history details of the subjects in the two groups in the hope of finding a significant correlation it engages in the preliminary stage of a process of hypothesis generation which, of course, belongs to the Context of Discovery.

Although an effort is made in case-control studies to address the issue of causal relevance (this is the purpose of contrasting the case and control groups), this type of study is obviously very far from implementing controls which are sufficiently systematic and rigorous to eliminate all but one (or a few) possible major aetiological factors. Indeed, if case-control studies did implement very strict controls they would tend to lose their heuristic function as potential generators of aetiological hypotheses and would become more akin to fully controlled experiments.

(ii) Cohort Studies

In cohort studies two or more groups (or cohorts) of people are identified and are then followed up over a period of time and compared. The aim is to see if any systematic differences appear between the groups after the original identification (which was on the basis of attributes defining group membership).

For example, we could look at a group of children all of whom had been orphaned in infancy and make assessments of their emotional development at intermittent periods until they reached adulthood. These results could then be compared with a group of similarly assessed non-orphaned children, also followed from infancy until adulthood. The aim is to draw conclusions either about the original difference or about the effects of experience in the intervening period (Backett and Robinson [1992] 204). Cohort studies are usually prospective (though Backett and Robinson say that if the effect has already taken place they can be seen as retrospective – ibid.). As with case-control studies, cohort studies are clearly not controlled in any strict sense.

(iii) Randomized Clinical Trials (RCT's)

These methods are widely regarded as coming closest to satisfying the criteria needed for discerning causal relevance along scientific lines in many areas of medicine. Consequently, in examining problems facing this format later on, we shall be considering difficulties for what is generally regarded as one of the best (or even the very best) clinical experimental format. The brief outline that follows (as well as additional discussion later in this chapter) is an *uncritical exposition* of the principles and statistical concepts involved. This is in accordance with the *classical* interpretation of such tests as presented in elementary textbooks (e.g. Miller [1984]). It is important to bear in mind that classical statistical inferential methods (due to statisticians such as Sir Ronald Fisher) have been strongly criticised – for example, by Bayesians. It would take us too far afield to become involved in the details of these controversies. However, for our present purposes it is perhaps relevant to note that because of these disputes internal to statistical methodology the problems of experimental and epidemiological tests are, if anything, even greater than conveyed here[35].

The basic clinical trials experiment involves a comparison between two groups, in which individuals are allocated to each of the groups by a random process. The aim of the randomization is to apportion to each group – as nearly as can be achieved – the same distribution of attributes which might conceivably influence the outcome and which cannot otherwise be controlled for. In addition, every effort is made to keep the conditions to which each of the groups is exposed

the same. The experimental intervention is then applied to one group, whereas the other group (the control group) receives a placebo. A placebo is an intervention which is intended to mimic all the non-specific effects which are concomitant with the application of the experimental variable.

After the experiment has been left to run for an appropriate length of time measurements are made of the values of the dependent variable for the individuals in the two groups (the independent variable being the experimental intervention, or its absence in the control group). The objective is to compare the distribution of values (of the dependent variable) obtained in the control group with that obtained in the experimental group. (One convenient way of doing the latter is to compare the means of the values for each of the two groups.) The two distributions will almost always be different. However, we are now confronted with a basic theoretical problem: how can we tell whether the two distributions differ because of: (i) extraneous variables which did not quite balance out during randomization (or because of other extraneous effects); or (ii) the presence vs. the absence of, specifically, the independent variable? It is the purpose of statistical testing to provide an answer to this problem. That is, it is the purpose of statistical testing to decide, on a probabilistic basis, whether the observed difference of outcomes between the two groups was due to: (i) purely the influence of extraneous variables (we may also call these 'chance' factors, so long as we remember that this is shorthand for the effects of factors not perfectly matched across the groups); or (ii) (at least in part) the influence of the experimental variable (Miller [1984] 49-50). The hypothesis that the difference between the two groups is due purely to extraneous or 'chance' factors is called the null hypothesis; whereas the hypothesis that it is due, at least in part, to the experimental variable is called the alternate hypothesis.

It needs to be emphasised that it is in the nature of statistical inference that a decision as to whether the null or the alternate hypothesis is true cannot be decided beyond all doubt, but only on a probabilistic basis. What is usually done is that a convention is adopted for accepting or rejecting a conclusion. This is usually done for the null hypothesis (i.e. the hypothesis that says that there is no significant difference between the groups) and it is set at a pre-agreed value (called the significance level, α): for example, $\alpha = 0.05$. The probability, p, that the null

hypothesis is true is then calculated. If $p \leq 0.05$ the null hypothesis is rejected (on the basis of the argument that a result with this value would only be expected 1 time in 20, or less), and the alternate hypothesis is accepted. That is, we conclude that there is a significant difference between the groups due to the influence of the independent variable. If $p > 0.05$ then we accept the null hypothesis that there is no significant difference between the experimental and control groups.

The actual mechanics of calculating the probability, p, of the null hypothesis by means of an appropriate statistical test need not concern us here. However, it will be helpful if we briefly consider the theory of samples and populations on which it is based. The theory on which statistical tests are based assumes that we can treat the samples with which we work (i.e. the control and experimental groups) as if they were samples that had been drawn at random from an ideal population. (N.B. A random sample is one in which the probability of selecting an individual is the same for all members of the population.) The mean value for each group can therefore be thought of as if it were the mean value of a sample, where each sample has been drawn from a parent population. In the words of Miller, the rationale underlying statistical testing can now be explained as follows:

> "Under the null (or chance) hypothesis the two groups of scores are samples from the *same* population and they differ because of sampling variability – the tendency for any two samples to vary slightly. On the other hand, the alternate hypothesis states that the two samples come from two different populations, and the sample means differ because the population means differ.
>
> The purpose of a statistical test is to find the probability that the chance hypothesis is true, i.e. the probability that the two samples came from a single population."
>
> (Miller [1984] 71; emphasis in original)

The important general point, for our purposes, is that the assessment of whether the observed difference between the control and experimental groups is significant or not (for a given, accepted level of significance) is based upon the probability that a difference between the means is either: (i) so small that it can be attributed to sampling variation in samples drawn from a single population (implying the null hypothesis); or (ii) is so large that it warrants the conclusion that the samples must have been drawn from different population distributions (implying the alternate hypothesis).

4.3 PROBLEMS FACING GROUP-COMPARISON EXPERIMENTAL TESTS

(a) Problems Related To Sample Size

From the standpoint of the statistical theory upon which clinical trials experiments are based, the size of the samples used is of extreme importance. The ability of the test to detect a true difference between the control and experimental groups is related to the size of these groups (Chassan [1979] 49-50). The larger the samples the greater the ability of the test to detect a genuine difference (if there is one), although beyond a certain size (for a given significance level and magnitude of the actual effect) the additional gain is negligible (op. cit. 50-51). An elementary, though basic, consideration is whether sufficiently large numbers of (suitably homogeneous) patients can in fact be obtained. Without the satisfaction of this fundamental practical requirement the possibility of an effective test does not even arise. If the samples used are too small the test will not be able to distinguish adequately between the variation occurring naturally between individual subjects and that occurring systematically across groups. Chassan says:

> "Perhaps the greatest difficulty of the extensive [i.e. group-comparison] model stems from the relatively small number of sufficiently homogeneous patients who can be brought together at any given time for the purpose of a specified clinical research project."
>
> (Chassan op. cit. 209. N.B. Kazdin [1980] 103-104 makes essentially the same point.)

Statistical theory provides methods for estimating the size of samples needed in order to detect a difference or effect of a given magnitude (if there is one) at a given level of significance (see e.g. Chassan op. cit. 50-52 and Johnson [1992] 51-53). However, regarding such estimations Chassan points out that:

> "There are, of course, many factors...which enter into such considerations. At this stage in the development of the art there can be no pat answers."
>
> (op. cit. 40)

Chassan also notes that:

> "A major difficulty...in applying [this] kind of information [i.e. regarding estimation of requisite sample size]...is that one never really knows beforehand just how large a true difference one may be dealing with as a guide toward a selection of the number of patients for a study."
>
> (op. cit. 50-52)

If the magnitude of the true difference is less than one judges then one might easily underestimate the size of the sample needed. On the other hand, if one errs on the side of caution by deliberately using a sample size that is larger than that dictated by one's best judgement of the actual difference, then one faces the practical problem of extra recruitment (and, as we shall see later, there are other costs of increasing sample size).

Given that there is at least some available theoretical guidance towards estimating effective sample size, it is regrettable that much research practice ("most" in Kazdin's view – see Kazdin [1980] 360, footnote 3), appears not to make proper use of it. Johnson says:

> "Many clinical trials and certainly most of those conducted in psychiatry give little if any consideration to the determination of the sample size at the design stage....Medical journals continue to publish useless studies with too few patients in each treatment group, without any indication of the implication of such sample sizes for the precision of the estimated differences between treatments. Thus one of the most important aspects of clinical trial design, namely the choice of sample size, is based not on some rational calculation but is left irresponsibly to be determined by the numbers of patients available at some clinic or hospital or who can be studied within some convenient set period."
>
> (Johnson [1992] 50-51; reference omitted)

Johnson (op. cit. 52) illustrates the sample sizes needed to detect a true difference between two groups at a significance level of $\alpha = 0.05$ when the power (i.e. $(1-\beta)$) of the test is 0.80. (N.B. The 'power' of a test is a measure the likelihood of its ability to detect a true difference, if there is one: i.e. of its ability to reject the null hypothesis when the latter is false. In the language of statistics the power equals $(1-\beta)$, where β is the probability of making a type II error. To commit a type II error is to accept the null hypothesis when it is false.) If there is a 20% difference in the effect between the two groups (when the absolute value of the response rate in the group with the weaker response is 30%), then the number of subjects per group needed to show up this difference is 95. To make these figures, which are based on mathematical theory, more concrete we can imagine a corresponding experiment in which we obtain, say, a 50% success rate in a group treated with a drug D and only a 30% success rate in the placebo control group (hence the difference between the groups is 20%). To be (at most) 80% sure that the test will show up this difference between the groups at a 0.05 significance level we

need at least 95 subjects per group. An effect equivalent to a 20% difference between the groups is substantial. Many effects are much weaker. Also, using a power of $(1-\beta) = 0.80$ and a significance level of $\alpha = 0.05$ are minimal. Yet even with these undemanding conditions a total of at least 190 subjects are needed. If the true difference of the effect is 10% (and still using a power of 0.80 and a significance level of 0.05) we would need at least 360 subjects *per group* to detect it (Johnson ibid.). Kazdin says: "Psychological experiments, for some reason, usually employ groups that are between 10 and 20 subjects" (op. cit. 360). One can only hope that matters have improved since Kazdin wrote that. Johnson comments with irony: "Compare these [i.e. the sample sizes dictated by statistical theory] with the actual sizes of published trials!" (ibid.).

To point to poorly designed or improperly conducted experiments is not, of course, to impugn experimental methodology *per se*. The present aim is to make more widely appreciated just how critical we need to be of experimental conclusions, given practical realities. Whereas experimenters can be blamed for poorly designed or executed experiments they cannot be blamed if sufficient numbers of suitable subjects simply cannot be found. As was emphasised in section 1.4(B), testing is always a practical matter, and if the conditions needed to test a hypothesis cannot be met in practice it will remain untested or inadequately tested. The often formidable practical problems facing recruitment cannot be waived aside as if they were inconsequential addenda to the methodology. The epistemic value of a method is only as good as its applicability and workability in practice.

The seriousness of the sample-size problem can be brought home by an example. R. D. G. Ferguson and J. G. Ferguson ([1994] 1201-1216) review the literature on attempts to evaluate the effectiveness of administering anti-blood-clotting agents intra-arterially to improve the condition of post-stroke patients. I assume that the situation they portray is not grossly atypical, if it is atypical at all. Furthermore, although the example is from organic medicine I do not think that, *prima facie*, one should expect matters to be any better for psychiatric conditions. Indeed, if anything, one would *prima facie* expect matters to be worse, given the presumably greater number of potentially extraneous variables affecting the latter. They say:

> "As noted by Freiman et al..., most of the trials performed are too small to answer the questions they were designed to address. Often, the preliminary estimate of patient availability is much higher than the actual number of eligible subjects willing to participate. It is sobering to remember that the NHLBI-funded Multiple Risk Factor Intervention Trial screened a total of 361662 men and, of these, randomized 12866. The time required for planning and recruitment was 44 months.... Despite the large sample size, Multiple Risk Factor Intervention Trial had inadequate power to test the primary hypothesis at the level of significance stipulated in the study protocol.... Sample size is a particular concern in the design of any acute stroke intervention trial. Mori et al... required 2 years and three participating centers to enroll 31 patients with acute carotid territory stroke in an intravenous alteplase study; del Zoppo et al... enrolled only 1 patient with a vertebrobasilar stroke during a 4-year recruitment period in a 17-center study; Casto et al... enrolled only 5 of 615 consecutive carotid territory stroke patients presenting within 5 hours of the onset of symptoms in their local intraarterial urokinase study. In short, the number of patients required to establish a biologically and statistically significant therapeutic effect exceeds the current potential for patient enrollment...."
>
> (Ferguson and Ferguson op. cit. 1211; references omitted)

Suppose, however, just for argument's sake, that extremely large numbers of suitable test subjects could be found. The benefit of using large numbers to detect a small effect then needs to be set alongside another consideration. As Edelson has pointed out:

> "as the groups become large, almost any difference between them becomes 'statistically significant'...even if such a difference is so small as to lack all practical or theoretical significance."
>
> (Edelson [1984] 65)

It does not, of course, follow that every small effect will be theoretically or practically insignificant. The small effect one (let us suppose) detects might have been theoretically predicted and have no plausible alternative explanation. In that case the large-sample study could be seen as an important research tool. However, it is worrying if, in moving to very large samples almost any effect becomes significant. Kazdin discusses this in a subsection of his [1980] entitled 'Potential Limitations [of statistical evaluation]'. He says:

> "Statistical evaluation is strongly emphasized in psychology; indeed, statistical significance often is regarded as the definitive test of whether the variables under investigation are important or worth pursuing. Yet statistical significance is a function of many different features of an experiment only *one* of which is whether there is a relation between the independent and dependent variables. [One of these features is sample size.]...
>
> ...Statistical significance is a direct function of sample size. That is, the larger the sample size, the smaller the group differences that are

> needed for statistical significance at a given level of confidence. Said another way, a given difference between two groups will gradually approach statistical significance as the size of the samples within each group is increased....
>
> ...if a large number of subjects is included, statistical significance is virtually assured. For example, investigations using hundreds or even thousands of subjects, often available when large-scale testing results are studied as in the military, have reported that statistical significance is virtually guaranteed no matter what the independent variable is to categorise the data into different groups...."
>
> (Kazdin [1980] 359-360; emphasis in original; references omitted)

If, with very large samples, virtually any independent variable can be shown to be statistically significant for a given dependent one, then the epistemic and theoretical value of the test risks becoming altogether empty. Furthermore, if there is an empirical dispute over whether X is statistically significant for Y, what maximum sample size should be allowed in the test which is carried out to decide the matter? A size just large enough so that the protagonists can win the argument, or one slightly smaller than that required to be statistically significant (so that their opponents can win)? It does not follow from this that the method is useless. There may be comparatively objective grounds for concluding that an effect of a certain magnitude is detectable, and that the absolute value of this is either too small to be theoretically significant, or otherwise large enough to be so. But there is also scope for possible differences of opinion and judgement (see also subsection 4.3(f), below). If there is no reliable consensual interpretation of the result its theoretical significance will be jeopardised.

(b) Non-Equivalence of Groups

This centrally important issue will be discussed separately in section 4.4 (pp. 129-132), so only brief mention need be made of it here. Essentially, the problem is that of ensuring that when two groups of subjects are compared in a clinical trial experiment, the initial distribution of attributes of potential relevance for the dependent variable is even or 'equivalent' between them. Without such equivalence the effectiveness of inferring a difference (if there is one) due to the experimental intervention will be weakened or jeopardised altogether. An uneven distribution of attributes potentially influencing the outcome obviously constitutes a potential source of bias.

(c) The Response of Subgroups

Comparison between two groups may be insensitive to how different subgroups within each main group respond to the independent variable. Subgroups may respond in ways which are different from – and even contrary to – the overall response. The response of a subgroup to the independent variable may be highly statistically significant, whereas the difference between the two main groups may not be statistically significant (or vice versa). Considered from the point of view of the significance test, it will not necessarily detect a genuine effect (due to the experimental intervention) if that effect is limited to a small enough subgroup of the sample used. Chassan refers to a hypothetical example provided by Hoffer and Osmand ([1961]) which illustrates this. Suppose nicotinamide is being used to treat toxic psychosis in the mid 1930s. This was at a time prior to it being known that pellagra – a disease which can result in skin lesions and insanity – was due specifically to nicotinamide deficiency. Chassan says:

> "[Suppose a] sample of 200 toxic schizophrenic-like psychoses are selected and randomized into two equal treatment groups. The assumption is made that the total group contains 20 pellagrins with few skin abnormalities, with 10 falling into each of the two treatment groups. We thus will have a total of 100 patients in each treatment group and 10 out of each 100, pellagrins. The patients in one treatment group will be given placebo capsules matching the nicotinamide to be administered to the other group. A reasonable assumption, for the purposes of illustrating the point, is then made that 9 of the 10 pellagrins in the treatment group will respond to the vitamin at an administered dosage of 250mg. per day, and 3 of the 10 pellagrins on placebo will respond to other conditions of the investigation, such as better food and other nonspecific hospital factors. It is further assumed that there are no improvements in the non-pellagrins in either treatment group. Considering the results taken as a whole, one comes up with 9 out of 100 cured on nicotinamide and 3 out of 100 cured in the placebo group, a result which by application of the chi-square test will provide a value which is clearly not significant, with the inference that the data of the trial gave no reliable evidence that nicotinamide is effective in the therapy of toxic psychosis in this group of patients which included pellagra psychoses, despite our present knowledge that this vitamin is an almost specific remedy for this disease."
>
> (Chassan [1979] 214-215)

Note that in this example the numbers of pellagrins in each main group is assumed to be the same – so the problem is not that these are unequally distributed. Rather, the marked positive effect of the experimental intervention on a subgroup is 'swamped' by its lack of effect in the remaining test subjects. Important information pertaining to this subgroup is lost by the group-comparison method.

Although this example is hypothetical, in any actual trial one cannot be sure that there are no subgroups which respond in a manner which differs from the overall result.

(d) Lack of Individualised Knowledge

Group-comparison designs do not, of course, directly provide information about individual test subjects[36]. This is not surprising since they operate with the variability that exists *between* subjects (Chassan [1979] 209). However, from a theoretical and practical point of view this does constitute a limitation since we would often like to know which individuals responded or failed to respond in a certain way, as well as the detailed nature of their response. Chassan says:

> "...among those who improved in the test-medication group [in a group-comparison experiment] it is in general statistically and logically impossible to distinguish the particular patients who improved because of the pharmacological properties of the test drug from those who improved in the manner of a placebo response. Thus in the extensive [i.e. group-comparison] approach...[one] does not know the particular [patients] who did thus improve....
>
> ...[In] view of the wide range of possibly relevant patient-characteristics that turn up in any clinical study this means that...[with group-comparison designs]...one cannot, in general, tie treatment effect to reasonably specific patient-characteristics and parameters."
>
> (Chassan op. cit. 217-218; emphasis omitted. See also op. cit. 235, 237)

In medicine in general what we seek is individualised knowledge, and in psychiatry this emphasis on the individual is, if anything, even stronger (see e.g. Kazdin [1980] 10-11, 106; Hersen and Barlow [1976] 1, 34; chapter 6 of this thesis). The importance of individualised knowledge is fairly evident. Psychiatric diagnostic categories (such as 'neurotic', 'schizophrenic' and 'depressive') admit of great variability, so that particular details may be of considerable importance in characterising a disorder or psychological problem in a given individual. In addition, response to therapeutic treatment may vary substantially from one individual to another. Chassan says:

> "Even in medical areas in which the illness is easily attributed to a specific invading microorganism, response to a given treatment against the same infection may vary from one patient to another. The situation in this

regard is ostensibly even more difficult in the treatment and in the evaluation of treatment in psychiatric illness."
(Chassan op. cit. 215)

Chassan concludes:

"...results from an extensive [i.e. group-comparison] design yield relatively little information for clinical practice concerning specific characteristics as a basis for the selection of a treatment for a given patient, or for a discriminating choice between, say, two treatments, both of which may be believed to be of some value within some broad diagnostic category."
(op. cit. 219-220)

Given the above, the question arises as to how group-comparison methods can be of service in providing the highly individualised knowledge that is often sought (I shall return to this problem in section 5.5(B)). If, as seems evident, group methods cannot directly test hypotheses about single individuals, then if the results which they do yield are to be relevant and applicable to the individual it needs to be shown how. In the absence of this connection being made explicit the criticism could be mounted against group-comparison methods that they by-pass a central aspect of what psychiatrists and psychotherapists are interested in – i.e. individualised knowledge.

What is likely to happen in practice is that the group or generic result will be used as part of a clinical judgement when a conclusion or decision about an individual is needed. However, a *judgement* does not provide the kind of objective and decisive result that experiments are widely regarded as capable of providing when operating at their best. For example, an important aspect of Eysenck's advocacy of experimental methods was that they are objective and decisive (see the quotation on page 22). These goals can hardly be well-served if, once the group-comparison experimental results are in, there is no alternative but to rely on personal judgement in applying them to the individual. (The general problem of judgement and objectivity in experimentation is elaborated in section 4.3(f).) On this point it is significant that the medical statistician Tony Johnson notes:

"The...process of deciding whether the results from a [clinical] trial can guide treatment of a particular patient is a less structured process [than methods used for entering patients into a trial and requires] clinical *intuition and judgement.*"
(Johnson [1992] 29; my emphasis)

(e) Problems Related To Sample And Population

If the conditions which theoretically justify inferences made from group-comparison designs are not met, what rational justification is there for relying on results based on such designs? One way in which the conditions which the theory demands may not be met (and, it seems, are regularly *not* met in practice) is in terms of the constitution of the samples and of the parent populations from which those samples are (supposedly) drawn. Chassan says:

> "Ideally, the application of statistics and probability theory for the purpose of drawing inferences...in one way or another depends upon the idea of sampling from hypothetically infinite populations, or processes. In theory one should draw a "random sample" from such a population, or distribution. Observations and responses in the sample and the statistical analysis based upon them should then lead to inferences about the specified population from which the sample was drawn."
>
> (Chassan op. cit. 220)

Chassan then goes on to point out how, often (or even typically), there are inadequacies in the manner in which either the sample or the population is defined:

> "Let us [consider] the question of drawing inferences from a set of results obtained from an extensive [i.e.group-comparison] design. What is the relation of sample to population in this model? First, it is noted that the group of patients selected from the available patient population for any psychiatric clinical investigation can only be defined in the broadest of terms. Thus it is typical to find a protocol, or set of criteria, for the selection of, say, 50 cases for a clinical trial to read as follows: diagnosis of schizophrenia, period of hospitalization greater than one year, age range 20 to 60, and so forth. ...Nevertheless, from any one such study to the next there will be considerable variation in patient-characteristics. Thus there are many different types of schizophrenia; there are many possible differences in the distribution of ages within a bracket such as 20 to 60 from one study involving 50 patients to another; there may likewise be considerable variation in period of hospitalization beyond one year; etc.
>
> Further, there are many possible highly individual characteristics and circumstances of patients of the sort which never enter into a set of criteria for the selection of patients but which may nevertheless be strongly related either negatively or positively to the effect of treatment. These will vary quantitatively and qualitatively from study to study with the same broad selection criteria. Also there will be differences in various aspects of the research setting which can have a bearing on the results.... Some such aspects can be apparent, while others, such as those pertaining to particular characteristics of the clinical observers, may not be."
>
> (Chassan op. cit. 221-222)

Chassan concludes:

> "Such extensive variations in possibly important variables relating to the outcome of treatment between patient samples from one study to another indicates that to each sample (or for that matter to any group of patients whether or not it is a study sample) there corresponds in some theoretical sense a population which if it has any meaning at all must be regarded as an *ad hoc* extension of the sample (or of the group). Though such a vague population may be given some theoretical definition in this way, from a practical point of view it can have very little meaning...."
> (Chassan op. cit. 222; emphasis in original)

We need to ask to what extent the mathematical-statistical conditions which theoretically underpin and justify group-comparison model inferences (insofar as those inferences are based on the assumption of random samples drawn from a parent population) are ever *met*. Do we have well-defined populations from which our samples (i.e. the control and experimental groups) are drawn, either theoretically or in practice? Is it true that our samples are (or are equivalent to) randomly drawn samples, as required by theory? Chassan's comments cast doubt on the presumption that these conditions are regularly met in the kinds of studies typically undertaken with psychiatric patients.

The properties used for entry into the trial group are the only ones used to define the hypothetical population which that group (conceived as a statistical sample) is treated as being drawn from. But, as Chassan points out (ibid.), it is very likely that there will be factors not included in this protocol definition of the sample and population which will affect the outcome. It follows that the population to which generalisation would be appropriate (on the basis of the *actual* sample properties) will be different from that to which generalisation will be made, given that the protocol criteria are used as the basis for defining the population. A similar problem arises if several studies to test the same hypothesis are conducted (Chassan ibid.). The theoretical population will have to be defined in terms of the properties of one of the samples used, but it is almost inevitable that the actual sample properties will vary from one study to another, even though the same hypothesis is being tested.

This leads Chassan to conclude that, typically, the sense of 'population' employed will be "an *ad hoc* extension of the sample" which "from a practical point of view...can have very little meaning"(ibid.). By this I understand

Chassan to mean that the concept of population employed will, in effect, be an *ad hoc* construction. To begin with, we do not have a population that is well-defined independently, or in advance, from which a random sample can be selected. Instead, the population is defined by the properties used in the protocol to define the entry of subjects into the trial. This is the case even though it is very likely that the sample used will include individual attributes that were not employed in the protocol definition (but which may influence the outcome). Peter Urbach, a Bayesian critic of classical statistical testing says:

> "The logic of [statistical] tests is essentially that of estimating population parameters, or testing differences between such parameters, and they are valid only if the samples were drawn independently and at random from a population. But which population? According to Bourke *et al.*, 'the null hypothesis that drug A has the same effect...as drug B refers, in a vague sense, to all patients similar to those included in the study.' Pocock characterizes the reference population as 'all patients with the disease eligible for the trial', which sounds less vague but is not really as it leaves 'eligible for the trial' undefined. To be sure, trial eligibility can be, and often is, precisely defined in the trial protocol. But such a definition cannot adequately characterize a reference population for the purposes of the normal statistical tests. For any reference population would need to include unknown sufferers in faraway places and those at present healthy or yet unborn who will contract the disease in the future, such people being the potential recipients of the treatment, should it prove effective in the trial; but you cannot draw random samples from hypothetical populations full of potential people.
>
> Those tests are, nevertheless, widely applied without the random sampling condition being met, which seems quite unjustified."
>
> (Urbach [1993] 1424; references omitted)

Criticisms similar to those expressed by Chassan and Urbach can be found in Miller ([1984] 160) and Johnson ([1992] 29).

Kazdin accepts that random selection from a population is a logical-statistical prerequisite for being able to generalise (back to that population) (Kazdin [1980] 123). He acknowledges that "there is an obvious problem in meeting this requirement...truly random selection beyond a very narrowly defined population is not possible" (op. cit. 123 and 125). However, he adds:

> "...there are different resolutions to this [problem]. Initially, one can assume that lawful relationships demonstrated in experiments with a given sample are likely to hold for other individuals who are similar. Results obtained with college students at one university would be expected in the usual case to hold for similar students at another university. Yet many of the variables that would be assumed to be unimportant across samples, including geography, could very well lead to differences in experimental results. Estimating the likelihood of the generality of findings

> depends upon the specific findings and the accumulation of knowledge about the factors that may influence a particular manipulation....
>
> ...As a general rule, the variables across which the results of a given intervention can be extended need to be determined empirically. After a phenomenon has been demonstrated, research needs to evaluate subject and environmental variables simultaneously to assess directly the generality of findings....
>
> (Kazdin op. cit. 124-125; reference to footnote omitted)

With regard to the claim that the suggestions offered constitute a "resolution" (ibid.) to the problem the following can be said:

First, it simply begs the question that a "lawful relationship" has been demonstrated by means of a statistical test if: (i) the successful working of the test was theoretically predicated on the drawing of random samples from a well-defined infinite (or very large) population and (ii) that requirement was not met.

Secondly, whereas Kazdin's account of how the scope of a particular result is ascertained by means of further (empirical) research and judgement probably matches common practice, it should be appreciated that, once again, this does not constitute a resolution of the sample-population problem discussed, nor a methodological vindication of significance-test methodology. It is always possible to muddle through somehow when we make inductive inferences, both with regard to reaching more general conclusions and more particular (i.e. individualised) ones. This is, surely, what we do most of the time in everyday life, as well as frequently in science. But what we need to ask in the present case is whether the generalisations we reach are correct and, more importantly, are so because the theoretical underpinnings of significance-test methodology are valid and were actually utilised in reaching them. Given the difficulties outlined, I think there must be doubt that any generalisations reached would be reached strictly and solely through the utilisation of the inductive machinery of the experimental test, as theoretically intended.

(f) Objectivity And Judgement In Experimentation

One respect in which experiments are seen to be advantageous over non-experimental reasoning is in their objectivity. There are, indeed, structural features of, at least, well-designed and well-executed experiments which in principle, lend greater objectivity to conclusions drawn from them than is the case with conclusions based on informal (i.e. non-experimental) reasoning by one or more individual. For

example, an effective statistical experiment could in principle, and fairly objectively, settle a dispute based on the pre-experimental judgement of individuals over whether, say, a particular fertiliser influenced yield in tomato plants (if the effect was marked). Similarly, to use a non-statistical example, the famous 'White-Spot' experiment (Holton and Brush [1985] 392; see also Worrall [1989]) was able objectively to settle prior differing opinions as to whether a spot of light would be observed in the centre of the shadow of an opaque disc against which a beam of light had been shone. However, as Newton-Smith has said:

> "A practising scientist is continually making judgements for which he can provide no justification beyond saying that that is how things strike him....
>
> ...The good experimenter...makes countless...decisions [without being able to provide an explicit justification at the time] in the design and execution of his experiments....
>
> ...happily we are rarely in the situation in which we have nothing to do but to follow our intuitions. But on occasion we have to, and one who ignores this will have a distorted picture of the scientific enterprise."
>
> (Newton-Smith [1981] 232-235)

The importance of judgement in designing an experiment and in interpreting the results is stressed by Chassan in his [1979] (op. cit. xvi).

In their [1973] Eysenck and Wilson promote specifically the objectivity of experimental tests as one of their characteristic methodological virtues (especially in opposition to informal inferences made on the basis of psychoanalytical training). For example, consider again the quotation from their book (see page 22 of this thesis) in which they say: "[experimental studies] alone can have the *objectivity*...to make proper judgement possible" (op. cit. 8; my emphasis). As acknowledged, in many cases experimental studies do have the potential to increase the objectivity of judgement. However, they may also leave us in a situation in which, in interpreting a result, so great a reliance has to be placed on individual judgement that it is questionable whether the conclusion deserves to be regarded as objective. Given this, the portrayal of experimentation as objective needs to be strongly qualified. Eysenck and Wilson do not appear to appreciate this in their book. Their [1973] comprises a selection of the best papers available at the time on the experimental testing of psychoanalytic hypotheses. Eysenck and Wilson regarded the predominant standard of papers they looked at as "appalling" (op. cit. xii) but felt that there were some (those selected for inclusion in their book) which

represented adequate testing of Freudian theory (ibid.). Having promoted experimental studies as alone being able to provide objectivity (op. cit. 8) (as well as being methodologically necessary for "those searching after truth" – ibid.), it is noteworthy that, after critically reviewing the papers in their collection, they invite the reader to reach his/her own conclusion about the results! In the 'Epilogue' of the above book they say:

> "Obviously the reader is free to come to his own conclusion on the basis of the evidence here presented, but we may perhaps be allowed to put forward a number of points which may be helpful in coming to a decision".
> (Eysenck and Wilson op. cit. 385)

There are, in fact, at least two other occasions on which Eysenck and Wilson invite the reader to come to his/her own decision about the experimental conclusions (op. cit. xiv and 13, respectively). But if this degree of laxity in interpreting an experimental result is permissible, to what extent is it justified to claim that the experimental result is "objective" or that the overall methodology (experimental performance plus interpretation) is so? Eysenck and Wilson cannot, of course, be blamed for the quality of the studies they reviewed. However, they clearly regarded that quality as adequate since, apart from saying so (op. cit. xii), they use at least some of the studies to reach the conclusion that some parts of Freudian theory are "disproved" by them (op. cit. 392).

In general, I think a more cautious portrayal of experimentation is deserved. Well-designed and well-executed experiments have the potential for utilising data and reaching conclusions in ways that are relatively objective and deserving of acceptance, the inevitability of judgement, interpretation and the theory-ladenness of data notwithstanding. However, much depends upon the field of research and the particular hypothesis under scrutiny. In some areas (e.g. the social and clinical sciences) there is perhaps greater scope for the input of interpretative bias. Kazdin says:

> "...statistical evaluation provides a criterion to separate probably veridical from possibly chance effects. Although subjectivity and bias can enter into the process of statistical evaluation, for example, in terms of the tests that are applied and the criteria for statistical significance, the goal of statistics is to provide a *relatively* bias-free and consistent method of interpreting results."
> (Kazdin [1980] 358-359; emphasis in original)

Kazdin's statement is, I think, more accurate than the impression of the general objectivity of experimentation created by the statement of Eysenck and Wilson quoted earlier. A "*relatively* bias-free and consistent method of interpreting results" (Kazdin ibid.) may be the best we have – but it is still allows for the input of subjectivity and bias.

4.4 GRÜNBAUM AND EYSENCK FAIL TO DRAW ADEQUATE ATTENTION TO THE PROBLEMS OF EXPERIMENTATION

(A) Grünbaum

Whereas Grünbaum is meticulous and searching in critically analysing psychoanalytic case-studies and in revealing the actual or potential inferential errors in them, no comparable undertaking is carried out by him for the 'scientific' alternative of experimental or epidemiological methods. This is in spite of the fact that the latter methods are prescribed by him as necessary for the "proper test of Freud's central hypotheses" (Grünbaum [1986] 228; see also Edelson [1986] 232). Giving prominent recognition to generic problems associated with experimental and epidemiological testing seems, however, to be essential if we are to be realistically appreciative of the epistemic standing of conclusions reached by these methods, and of what we can hope to achieve by using them in psychological research.

Grünbaum *does* critically discuss some issues connected with experimentation. For example, in his [1984] (pp. 202-205) he criticises the otherwise significant experimental investigations by Motley into verbal slips for not being genuine tests of psychoanalytic theory[37]. Critical evaluations of various experimental studies (in a broad sense) are also provided by Grünbaum in his [1986] (pp. 269-270). In the latter case, as with the Motley studies, Grünbaum's principal aim is to show why the results fail to support Freudian theory. This is different from the point I am stressing here: namely, that systematic and careful attention needs to be given to how experiments can, and sometimes do, fail epistemically. *Part* of the latter is undoubtedly implicit in any examination of an individual experimental study which aims to show what conclusions can or cannot validly be drawn from it. However, my point is that a more general account of experimental limitations and inadequacies also needs to be presented.

Grünbaum also provides a critical examination of several strategies (e.g. causal modelling and statistical controls) and single-subject designs (time-series, ABA, and multiple baseline) (Grünbaum [1993] 233-242). This is in connection with Edelson's espousal of these strategies and designs for testing hypotheses in the psychoanalytical situation. Grünbaum argues that: "not even one of the single-subject designs enumerated by Edelson sustains his thesis that psychoanalytic *causal* hypotheses can be *cogently* tested in the treatment setting"

(Grünbaum [1993] 234; emphases in original). For example, the time-series design applied to testing whether a single intervention has a lasting therapeutic effect on a given disorder is criticised because it does not have the capacity to rule out the rival hypothesis of 'spontaneous remission' in the eventuality of the patient's improvement (op. cit. 235). But even if a favourable therapeutic outcome were attributed to the treatment intervention, the time-series design does not allow us to discriminate whether that outcome is due to some specific factor in the treatment (e.g. insight) or was a result of placebo effect (ibid.). In a comparable way, Grünbaum systematically criticises the other single-subject designs advocated by Edelson (Grünbaum op. cit.). Clearly, this *does* amount to a negative appraisal of the ability of these quasi-experimental methods to test effectively the kinds of hypotheses at issue.

A central aim of Grünbaum's criticism of the above-mentioned single-subject designs is to compare their liabilities with the advantages of group-comparison designs (Grünbaum [1993] 238-239, 242). But although Grünbaum expresses a methodological preference for group-comparison designs (see page 103 of this thesis), it is his failure to treat their potential liabilities and limitations with anything like the acumen with which he deals with the psychoanalysts' chosen methods that make him vulnerable to the charge of having double standards. My claim is *not* that informal reasoning or the use of single-subject designs are 'as good', epistemically, as group comparison designs; nor that because group-comparison designs have defects there is nothing to choose between them and the former; nor that reliable results cannot be achieved by group-comparison designs. It is, rather, that Grünbaum is negligent with regard to potential problems facing the designs he favours and that this results in him not being even-handed.

In his [1984] Edelson lists various problems facing group-comparison designs. It is quite likely that Edelson utilised Chassan's [1979] critique (as Chassan's work is referenced by him), and the reader may recognise some of the problems discussed by Chassan that were covered in section 4.3. Some of Edelson's points are as follows:

> "...group-comparison studies as carried out are often riddled with what are – given the stakes for knowledge and practice – breathtakingly nontrivial defects, which given the realities of clinical research are largely unavoidable.

> It is not easy to obtain truly random samples in clinical practice and settings. But then the effects of extraneous variables cannot be assumed to be randomized and equivalent in the compared groups, and these groups cannot be assumed to be similar with respect to all relevant variables other than the explanatory variable(s). Nor can these nonrandom samples be assumed to be representative of the populations from which they are presumably drawn. Therefore, the investigator lacks justification to generalise his findings to these populations....
>
> ...More important, the comparison of groups in terms of the mean or average effect of different factors, conditions, or treatments – which is the very heart of the group-comparison method – tends to obscure the nature of these effects. An average outcome is no single individual's outcome. ...Generalization from such a [group-comparison] study to a particular individual patient in clinical practice is, therefore, not possible....
>
> ...One or two members of a group may swing a group so far one way, despite the positions of the other members, that a group which would not otherwise be judged to differ in any significant way from another group is judged to do so."
>
> (Edelson [1984] 63-65)

Grünbaum was certainly aware of the problems raised by Edelson, since he refers explicitly to the passage in which they occur (Grünbaum [1993] 243). However, it should be noted that Grünbaum truncates Edelson's list, misrepresenting it as only a single problem (the first one mentioned by Edelson, concerning the non-equivalence of groups due to the inability to obtain truly random samples). Grünbaum makes no mention of, for example, the problem of what population the result should be generalised to (Edelson op. cit. 63-65), or of what parent population the samples are supposed to be drawn from (op. cit. 64); or the problem of how to apply the result from a group-comparison study to a single individual (op. cit. 64).

Worst of all, in a manner that is atypical of Grünbaum's normally exemplary standard of analysis, he descends into a form of *tu quoque* argument against Edelson in which he seeks to excuse himself from having epistemically to justify his preferred methodology by appealing to Edelson's error in advocating the method of 'systematic replication' in single-subject research. By 'systematic replication' Edelson had meant a strategy by which the result of a single-subject study could potentially be generalised to other subjects by finding examples in which the test subject differs from the original in only one property at a time (Edelson op. cit. 66). Grünbaum is justified in criticising Edelson's naivety in expecting that individuals or settings could be found which differed from one another in only a single property, and that the problem of extraneous variables would not confound the intended inference (Grünbaum [1993] 242-243). However, Grünbaum then says, surprisingly:

> "Yet, I submit, what is sauce for the goose is sauce for the gander. Any methodological dispensation that is granted to single-subject replication ought to be likewise explicitly accorded to group comparisons. What then entitles [Edelson] to deplore that I am beholden to the group-comparison method for psychoanalytic research?"
> (Grünbaum op. cit. 243)

Thus, Grünbaum utilises Edelson's reliance upon the (in practice unattainable) condition of ruling out the possible effects of extraneous variables in the method of systematic replication as an *excuse* for it not being incumbent on him to provide a satisfactory solution to their potentially confounding influence in group-comparison studies. However, Grünbaum surely ought to recognise that citing the failure of another person's advocated methodology does absolutely nothing to rectify the potential liabilities in one's own or to provide, in general, a justification for one's own methodology. This is perhaps especially so if the basis for criticising the other's methodology is logically similar to the defect in one's own (because in damning the other one thereby damns oneself). Thus, rather than Grünbaum being allowed to have a "methodological dispensation" (Grünbaum ibid.) to overlook, partly or wholly, the problem of extraneous causes just because Edelson has done so in the case of systematic replication, I would argue that neither author be allowed to have any dispensation at all for any strategy which carries a serious risk of error. Moreover, it is particularly regrettable that Grünbaum should have now implicitly abandoned the cause which he initially championed: namely, that of meticulously searching out and exposing possible sources of error in causal relevance inferences. After all, that was an issue which he made his own and which he thoroughly castigated the psychoanalysts for failing to address in their clinical case-study methodology. But then these same standards of care and rigour should be made to apply even when the causal relevance inferences are being made through the use of group-comparison studies.

Grünbaum does say:

> "...Edelson fails to mention Kazdin's discussion of standard strategies used to overcome the defects of comparative group studies."
> (Grünbaum [1993] 243)

However, this reference to alleged solutions to those problems is inadequate. Grünbaum does not specify which problem (or problems) he believes is

"overcome" (ibid.), or how. There certainly are standard strategies which attempt to improve upon at least some of the problems raised by Edelson, Chassan and others. However, there appears to be no unanimous agreement amongst specialists that they have all been successfully "overcome", as Grünbaum's remark (ibid.) would have us believe.

Let us consider as an example the problem which Grünbaum does explicitly recognise (Grünbaum [1993] 243): namely, that of whether the potentially confounding effects of extraneous variables can be eliminated when two groups are compared. Kazdin recognises that:

> "Randomly assigning subjects [to groups] *can* produce groups that differ on all sorts of measures...[especially] when sample sizes are small and when there are extreme scores in the sample."
> (Kazdin [1980] 126; emphasis in original)

Kazdin goes on to say that with large numbers of subjects random assignment: "is likely to produce groups that do not differ significantly on characteristics relevant to treatments" and that prior to experimental manipulation the groups can be assessed to see whether they differ with respect to such factors as age, IQ, years of institutionalisation etc. (op. cit. 126-127). However, Kazdin cautions:

> "Of course, such results do not establish absolutely that the groups are equivalent because groups still may differ on some variable, *relevant* or irrelevant, that the investigator did not assess."
> (Kazdin op. cit. 127; my emphasis)

Moreover, it is not clear what should be done if the groups were found *not* to be equivalent at the pre-treatment assessment. Should the experiment be abandoned? Or should the experimenters try to balance out the groups by moving subjects across from one group to the other (this would violate the conditions needed to satisfy randomization – see Urbach [1985] 265-266)?

A technique which can be used to avoid leaving to chance whether a given variable (of potential relevance to the dependent measure) is equivalently distributed amongst groups is 'matching' (Kazdin op.cit. 127). Matching involves "grouping subjects together on the basis of their similarity on a particular variable or set of variables" (Kazdin ibid.), and then randomly assigning those individuals to the

groups. However, unless the magnitude of the 'matched' variable is the same for each subject in the matched sub-group the groups will not be exactly equivalent. Further, there will be no guarantee, if matching is carried out with respect to one variable, A, that other variables (e.g. B, C, D etc.) which are relevant for the outcome, will be equivalently distributed amongst the groups. Suppose, for example, that we can find two test subjects, John and Peter, who are virtually identical on their ratings for variable A. They therefore form a matched pair with respect to that variable and can be randomly assigned, one to each group. But suppose that John, apart from being an A also has attributes B and C (but not D), whereas Peter is D but not B or C. The groups will then be non-equivalent for B, C, and D, although matched for A.

Kazdin concludes that:

> "Matching followed by random assignment is an excellent way to ensure equivalence of groups on a measure that relates to the dependent variable."
> (Kazdin [1980] 157-158)

This is in spite of the fact that he earlier acknowledged that:

> "...one cannot realistically expect to implement an experiment where all conditions are the same except for the independent variable."
> (op. cit. 52)

Chassan is less optimistic than Kazdin about attaining equivalent groups both when there is straightforward (or 'complete') randomization, and when matching (or to use Chassan's terminology, "pairing or grouping") is employed prior to randomization. With regard to the completely randomized design Chassan says:

> "Theoretically, the justification for the use of a completely randomized design lies in the assumption that the experimental units – in clinical trials based on the extensive [i.e. group-comparison] model, the patients – can more or less be considered relatively homogeneous with respect to the outcome variables. When the experimental units are not homogeneous a lack of pairing or grouping at best results in a relatively weak design.... At worst, a heterogeneity in relevant attributes among the experimental units can contribute to overt or hidden biases to the extent that erroneous conclusions can be drawn from the results of such a study."
> (Chassan [1979] 148)

After considering a suggestion by Radhakrishna and Sutherland ([1962]) that analysis of covariance and standardisation be used as possible ways of dealing with the problem of non-homogeneity, Chassan concludes:

> "These [techniques], [although useful and valid for handling overt biases, especially with a large number of patients and a small number of biasing characteristics,]...do not suffice for the handling of hidden bias in the variability between patients, a consideration of particular importance in the design of psychiatric trials. There appears to be no solution to this problem within the framework of the extensive [i.e. group-comparison] model of design...."
>
> (Chassan op. cit. 148-149)

As for matching (i.e. grouping or pairing) prior to randomization, Chassan is scarcely less critical:

> "...it is in part the relatively large number of patient-characteristics itself in relation to the comparatively small number of patients in any one study that makes grouping or pairing of patients prior to the allocation of treatment a necessarily incomplete and generally unrewarding procedure. There are so many variables with which one can logically initiate the grouping or pairing process that one often hardly knows where to begin. And once started, the point at which one must end soon becomes evident, not for a lack of variables which ought to be taken into account, but because it often becomes evident that the number of combinations of patient-attributes that may have a bearing on prognosis achieves astronomical proportions in relation to the number of patients that can be available in a given study. One is then faced with a dilemma: the larger the number of patients in a study, the greater is the likelihood for increasing the variation in these attributes!"
>
> (Chassan op. cit. 149)

There is also a practical constraint on matching. Evans says:

> "In many trials there is no attempt to identify prognostic factors, and simple randomization, often constrained to achieve approximately equal numbers of patients on each treatment, is used to form the groups. [It may be thought that both the views of the classical statistician R. J. Fisher] and common sense dictate that experimental groups must be deliberately matched on all known prognostic factors. This is where practical considerations overtake any philosophical niceties. Patients do not arrive for treatment in any matched order, and one cannot delay their treatment until a patient with similar values for all prognostic factors turns up, or even allocate them provisionally to one group and allocate the next patient with those same factors to the other group....
>
> ...It is [in practice] only very rarely possible to obtain matched pairs of patients in clinical trials, except in those (such as crossover trials) where treatments are compared 'within patients' by giving more than one treatment to each patient."
>
> (Evans [1993] 1433-1434)

A further critical comment is made by Miller:

> "A major drawback of the matched subjects design is the difficulty of knowing which subject variables should form the basis of the matching. And even if we can identify the most influential variables, it is often a difficult and time consuming matter to actually recruit pairs of subjects whose scores on these variables are closely matched. Consequently this design is not met very frequently outside the literature on twin studies."
> (Miller [1984] 13)

Overall, given these critical reservations, Kazdin's statement that "Matching followed by random assignment is an excellent way to *ensure* equivalence of groups" (ibid.; my emphasis) must be called into doubt. Similarly, the reference made by Grünbaum to "Kazdin's discussion of standard strategies used to *overcome* the defects of comparative group studies" (Grünbaum op. cit.; my emphasis) is unconvincing, at least with regard to the problem of group equivalence. Of the problems facing group-comparison designs that Edelson raised (op. cit.) the latter was, moreover, the only one recognised by Grünbaum (see page 127 of this thesis).

(B) Eysenck

In those of his works covered here which deal with the critical appraisal of psychoanalysis (Eysenck [1953], [1985]; Eysenck and Wilson [1973]) Eysenck does indicate – in a manner which is more explicit than Grünbaum – that there are fundamental difficulties associated with experimental inference. For example, Eysenck says:

> "...when a special experiment is carefully planned to test the adequacy of a given hypothesis there often arise almost insuperable difficulties in ruling out irrelevant factors [i.e. extraneous variables], and in isolating the desired effect; in clinical work such isolation is all but impossible."
> (Eysenck [1953] 229)

And:

> "It is possible to hold that an experiment is performed only when all conditions are rigidly controlled, with the exception of the independent

> variable, which is systematically varied to permit measurement of concomitant changes in the dependent variable. It would be safe to say that...such experiments simply do not exist in psychology."
> (Eysenck and Wilson [1973] xiii-xiv)

In addition to the above acknowledgements of the difficulties inherent in experimental inference there is a passage in which (for the case of clinical trials) Eysenck says that he will "deal" with the difficulties (Eysenck [1985] 49). However, as I shall now argue, Eysenck's treatment of those difficulties is grossly inadequate. The relevant passage occurs in a section in which Eysenck discusses the problem of how to evaluate reliably the therapeutic effectiveness of psychoanalytic treatment. Eysenck begins by saying:

> "It is a curious feature of psychoanalysis that until relatively recently very little attempt was made to demonstrate its effectiveness. Right from the beginning Freud himself opposed the usual medical practice of instituting clinical trials to assess the efficacy of a new method of therapy, and his followers have slavishly adopted the same stand."
> (Eysenck [1985] 46)

It should be noted as an incidental point that Eysenck's remark is historically inaccurate. "[The]...medical practice of instituting clinical trials to assess the efficacy of a new method of therapy" was *not* standard or "usual" (Eysenck ibid.) until after the time of Freud's death in 1939. Johnson, for example, says: "Since its inception in the *1940s*, the randomised clinical trial has become the principal method of comparing the efficacy of all forms of medical treatment and care..." (Johnson [1992] 24; my emphasis). Consequently, Eysenck criticises Freud for not employing a method that was not standard procedure in his lifetime even though, of course, it is uncertain whether Freud would have embraced it had it been so. Eysenck then goes on to say:

> "One reason sometimes given by psychoanalysts for not conducting a clinical trial, with an experimental and a control group, and a long-term follow-up, is the difficulty of such an undertaking. There is no doubt about the difficulties involved, and we shall deal with these presently...."
> (Eysenck op. cit. 49)

I now want to examine precisely what Eysenck takes to be the difficulties involved with clinical trials and how he "deals" with them.

(i) Eysenck says:

> "Of the difficulties that arise, the most important is perhaps the question of the criterion to be accepted for improvement or cure [of a neurosis]."
> (op. cit. 50).

The question that Eysenck raises has a bearing both on the theory of neurosis and on the methodological problem of selecting an appropriate dependent variable and measuring it. Eysenck's discussion (op. cit.) deals with the first of these. He contrasts (what he considers to be) the advantages of the Behaviourist approach to neurosis (which limits itself to removing such 'observable' symptoms as phobias, depression, anxiety attacks, obsessions and compulsions) with the psychoanalytical approach (which concerns itself with the underlying 'complex'). Eysenck does not discuss the second issue, relating to the methodological problem of what would be an appropriate measure of the theoretical construct of interest. The latter is a fairly standard topic in research methodology since the dependent variable must represent the concept(s) of theoretical interest, and yet must be capable of empirical evaluation (see e.g. Kazdin [1980] chapter 9; Freeman and Tyrer [1992] chapters 6, 13 and 14). However, in spite of its general methodological importance, this issue is not a specific problem for *clinical trials* inference (it would be relevant for any study in which a dependent variable was being assessed). Not only does Eysenck not discuss the methodological problems concerned with the selection and evaluation of the dependent variable (he discusses instead issues related to psychological theory) but, even if he had, it would not have been a specific problem for clinical trials inference. Given that his stated intention is to deal with the latter his comments are very largely, if not completely, irrelevant.

(ii) The second of the difficulties for clinical trials which Eysenck says he is going to consider is "the question of the make-up of the experimental and control groups" (op. cit. 51), which he also describes as "the selection problem" (ibid.). Although Eysenck states this as his aim he seems to digress straight away by informing us instead about how psychoanalysts select for treatment only a very restricted range of clients:

> "Psychoanalysts are very definite that their treatment is only suitable for a very small percentage of neurotic patients; they are very careful in their criteria for selection. In preference a patient should be young, well-educated, not too seriously ill, and reasonably well-off...."
> (Eysenck op. cit. 51)

Eysenck then elaborates on the unrepresentativeness involved:

> "...in one typical study, 64 per cent of patients undergoing analysis had received postgraduate education (as compared with no more than 2 per cent or 3 per cent of the general population), 72 per cent were in professional and academic work, and approximately half of all the cases were 'engaged in work related to psychiatry and psychoanalysis'."
> (ibid.)

Let us assume that Eysenck's depiction of the kinds of persons entered for psychoanalysis is accurate. This information would be important if what was at stake was an actual trial to assess the effectiveness of psychoanalysis amongst the population at large, since the participating clients would be unrepresentative. However, no trial is being conducted. Moreover, Eysenck is perfectly aware of this since only a few pages earlier he had said:

> "...Freud...opposed...clinical trials...and his followers have slavishly adopted the same stand."
> (op. cit. 46)

And:

> "...until successful [clinical] trials are completed, psychoanalysts have no right to make any claims. The fact [is] that they have hitherto completely shunned this duty [i.e. the duty of conducting clinical trials of psychoanalytic treatment]...."
> (op. cit. 49)

Consequently, the relevance of Eysenck's description of the characteristics of psychoanalytic clients for the internal methodological problems of clinical trials is utterly obscure, or even missing. By his own agenda Eysenck is supposed to be dealing with the inherent problems of clinical trials. If Eysenck does have a point it would be that, possibly, psychoanalytic clients would be better fitted for recovery than the general population because of their being better educated, having higher socio-economic status etc.. But this is not an internal problem of clinical trial inference.

In fact, Eysenck does not even explain what "the question [i.e. problem] of the make-up of the experimental and control groups" (ibid.) is, and so this is left to the surmise of the reader. My own interpretation of what could be subsumed by it includes: (i) problems relating to the selection of samples which are representative of the population; and (ii) problems relating to whether the experimental and control groups are 'equivalent' prior to the experimental intervention. There is, however, a danger of reading into Eysenck's discussion topics which he did not have in mind – a risk which stems from his failure to set out clearly the problem he wishes to tackle and then deal with it. Especially regrettable is the wasted opportunity of not addressing important problems inherent in clinical trials methodology.

Eysenck goes on to consider three further difficulties:

(iii) Eysenck says:

> "Another difficulty is the control group. If denied treatment, are they [i.e. persons in the control group of a clinical trial in which the treatment group receives psychoanalysis] not likely to seek help elsewhere – either by going to a general physician or a priest, or by discussing their problems with friends or members of the family...."
> (op. cit. 51)

The difficulty which Eysenck is alluding to is that of ensuring that control group members are not influenced by background or extraneous factors which are relevant for the effect being measured. In the present case, discussing personal problems with friends etc. might serve as a form of 'informal psychotherapy', thereby perhaps showing the experimental group (receiving psychoanalysis) to be less effective by comparison than it otherwise would be. Whereas Eysenck raises this problem he does not discuss it or attempt to provide any solutions. Indeed, Eysenck's treatment consists of three sentences. This includes the two sentences from which the above quotation is taken, plus a third in which Eysenck asks the question "how can we prevent members of our control group from making use of such facilities [i.e. airing their problems with friends etc.]?" (op. cit. 52). No attempt is made to answer the question.

(iv) The fourth difficulty concerns 'placebo effect'. By the latter is meant a therapeutically beneficial effect brought about by non-specific aspects of psychotherapy (for example, sympathetic attention, or obtaining advice from a medical authority-figure), rather than because of any specific component of psychoanalytic theory or therapy. Eysenck rightly points out that in order to carry out a clinical trial on the effectiveness of psychoanalysis it would be necessary to administer a placebo treatment to the control group which makes available to its members the same kinds of non-specific effects that the experimental group (undergoing psychoanalysis) receives. This is so that non-specific therapeutic influences would be equal in the two groups and any subsequent differences in therapeutic outcome could be attributed, *ceteris paribus*, to the specific effects of psychoanalysis.

Eysenck then points out the difficulty involved in designing an effective placebo treatment:

> "...a placebo control group is really essential if the trial is to be taken very seriously. However, it is, of course, difficult to design a treatment which fulfils the function of the placebo in not containing any of the specific parts of the experimental treatment, but is also acceptable as meaningful to the patients involved! It is not impossible to devise such placebo treatments, but it obviously needs a good deal of thought and experience."
>
> (op. cit. 52)

In this case Eysenck has identified a genuine difficulty facing the designers of clinical trials who seek to test the effectiveness of different treatment modalities. However, no specific way of solving the difficulty is proposed.

(v) Eysenck acknowledges that "there are many other difficulties [facing clinical trials]" (op. cit. 52) but specifically mentions and discusses only one: namely, the ethical problem of "how can we really justify the withholding of a successful treatment from the control group of patients, simply because of our scientific curiosity?" (op. cit. 53). Eysenck responds by arguing that prior to carrying out a clinical trial we have no reason to believe that a given treatment is effective, even if it has been widely employed on the basis of the assumption of its effectiveness. He concludes:

> "Once a particular method of treatment has been found efficacious by clinical trial, it may be unethical to deny it to patients; while it is still questionable whether it has any effect at all, or arguable that it may instead have a negative effect, i.e. make the patient worse, as has been suggested for psychoanalysis, no ethical problem arises."
> (op. cit. 53)

Here again, Eysenck identifies a genuine problem associated with clinical trials practice, but his attempted resolution is overly simple and inadequate.

First of all, his claim about lack of knowledge prior to well-conducted clinical trials is not wholly accurate. Sometimes the anecdotal evidence for the efficacy or otherwise of a particular treatment is sufficiently strong to warrant its use (as compared with no treatment), even though Eysenck is correct in implying that the thorough investigation of the treatment's efficacy would be preferable. For example, St. John's wort (Hypericum perforatum), a traditional herbal remedy for depression and anxiety, has recently been found to be "as effective as standard antidepressant therapy, according to a major research trial" ('BBC News Online – Health' for 31.8.2000: http://news.bbc.co.uk). The 'folk-knowledge' of this herb's therapeutic efficacy seems therefore to have been borne out by the recent trial (headed by Dr. Helmut Woelk of the University of Giessen, Germany). There are cases in which there is a genuine ethical dilemma about withholding a treatment that seems to have some efficacy, but has not been thoroughly tested in clinical trials (I am not, however, suggesting that Freudian psychoanalysis is one of these).

The weakness in Eysenck's argument can be brought home by an extreme example (as an analogy). To the best of my knowledge, bread has never been evaluated by means of clinical trials for its effect on nourishment in humans (let us assume it has not been). Bread's positive causal relevance for nourishment has nevertheless been informally established through 'folk-reasoning' over thousands of years. Would Eysenck wish to withhold supplying bread to, say, famine victims on the grounds that it has not "been found efficacious by clinical trial" (ibid.), and would he argue that until it were so tested "no ethical problem arises" (ibid.)?

Also, Eysenck does not mention the type of case in which we have a treatment with clinical trial proven efficacy, but comparison is being sought against a possibly even more effective rival treatment, but where the margin of gain might be small. Would it be ethical to conduct a trial using one group on the new treatment

and one on the old, given that the new treatment might have side-effects outweighing any possible benefits (and let us assume that the old treatment has acceptable side-effects)? In this case one cannot truthfully claim that there was no available effective treatment, and one cannot be confident that the risk to patients receiving the new treatment is acceptable. There is, therefore, an ethical dilemma about whether or not to test the new treatment by this method.

Overall, ethical decisions about testing new treatments are not as simple and clear cut as Eysenck makes them out to be.

In conclusion, of the five 'difficulties' facing clinical trials which Eysenck discusses only (iii), (iv) and (v) are relevant and, of these, a solution (or, rather, rationalisation) is provided for only one of them (i.e. (v)). More importantly, Eysenck does not include in his list any of the major problems raised by Edelson, Chassan and others that were discussed earlier in this chapter.

4.5 SOME PROBLEMS OF TESTING - AN EXAMPLE

To give an idea of some of the difficulties facing the testing of some psychoanalytical hypotheses by group-comparison methods I shall now consider an example used by Grünbaum. The example is taken from Freud's famous case of the 'Rat-Man', Paul Lorenz (Freud [1909]). Freud had hypothesised that Lorenz's adult psychopathological symptoms of obsessional neurosis were due to his having been punished as an infant by his father for masturbating, leading to the repression of his infantile sexual inclinations. I shall treat the core of Freud's claim as being about the repression of infantile sexual inclinations because of actual parental punishment for an act of genital stimulation. Variants of this hypothesis could, of course, be formulated. For example, cases in which repression occurred in infancy but there was no external punitive act (e.g. the repression was due to a feared or imagined punishment rather than a real one); or in which there was punishment for an act with a sexual theme, but it was not specifically for genital stimulation.

Grünbaum criticises Glymour's [1974] and [1980] reconstruction of the Paul Lorenz case in which Glymour had claimed that Freud's theory "was strong enough to be tested on the couch [i.e. during the interview sessions]" (Glymour [1974] 304). Grünbaum objects that even if Freud could have correctly ascertained (during the course of the clinical sessions) that, in his infancy, Lorenz had been punished by his father for masturbating "this much would be quite insufficient to support Freud's etiologic hypothesis that repressed precocious sexual activity is *causally relevant* to adult obsessional neurosis" (Grünbaum [1984] 251-252; emphasis in original).

Grünbaum's point is, of course, a crucially important one. Suppose Lorenz had been punished for the reason Freud hypothesised (and, according to Grünbaum, there was no direct evidence for this – Grünbaum op. cit. 253). It was also a fact that Lorenz suffered from adult neurotic symptoms. But why should it be supposed that the former had any causal influence on the latter? It is, after all, the claim that there is a *causal connection* between the variables that constitutes Freud's hypothesis. But what if Lorenz would have developed the same symptoms even if he had never had any infantile sexual inclinations; or if he had had the inclinations but had not repressed them; or if he had repressed the inclinations because of parental punishment but this had not made any difference to the appearance of the adult

symptoms? The crucial point, if the hypothesis is to be sustained, is that a causal relationship must be demonstrated between punishment for expressing (or repression of) infantile sexual inclinations and adult symptoms. In the absence of a satisfactory demonstration we cannot be sure that a *post hoc ergo propter hoc* fallacy has not been committed. Grünbaum says:

> "Let "N" (neurosis) denote a psychoneurosis such as the syndrome of obsessional neurosis, and let "P" (pathogen) denote the kind of sex-related antecedent event that Freud postulated to be the specific cause of N. ...To support Freud's etiologic hypothesis that P is causally necessary [or causally relevant] for N, evidence must be produced to show that being a P *makes a difference* to being an N. But such causal relevance is *not* attested by *mere* instances of N that were Ps, i.e., by patients who are both Ps and Ns."
>
> (Grünbaum [1984] 253; emphases in original)

However, whereas Grünbaum is entirely justified in emphasising that the demonstration of a causal connection between two events (or two classes of events) requires much more than demonstrating their co-occurrence, he is insufficiently critical of the method he proposes for achieving this end in the example at hand. Grünbaum goes on to say:

> "Now contrast the stated epistemic liabilities of the retrospective psychoanalytic inference that a given adult patient was or was not a P during his early childhood with the assets of *prospective* controlled inquiry: a *present* determination would be made, under suitably supervised conditions, whether children in the experimental and control groups are Ps and non-Ps, respectively; again, during long-term follow-ups, later findings as to N or non-N would be gathered and would pertain to the then state."
>
> (Grünbaum op. cit. 258-259; emphases in original)

In other words, Grünbaum recommends using a prospective group-comparison method for testing the aetiological hypothesis. I shall now argue that this proposal is actually impaired by a variety of problems which Grünbaum does not appear to recognise. Precisely which group-comparison format shall we use? Of those which have been discussed in this thesis (see 4.2) the case-control study must be excluded, since it is a retrospective design, and Grünbaum is insistent that we should use a prospective method. This leaves either cohort studies or RCTs. Let us consider these in turn:

(i) Cohort studies.

If we use a cohort study it would, of course, be possible in principle to form two groups (e.g. infants who had been punished by a parent for masturbating and had repressed their sexual inclinations because of the punishment; and those who had not been reprimanded in any way for sexual expression and were sexually unrepressed). In practice, however, it would be very difficult to do this. There are problems about what exactly is meant by the concept of repression, how we could tell whether repression had occurred, and whether it was induced by punishment or was self-induced. But let us leave aside these difficulties since they are quite general and do not pertain specifically to cohort or other group-comparison designs. What cannot be overlooked is that infants are unlikely to come forward with complaints about parental impositions on their sexual expression (assuming Freud's theory about the ubiquity of infantile sexual inclinations to be correct). Also, parents may be unwilling to acknowledge to themselves that their actions might damage their children, let alone volunteer participation in a study designed to assess the long-term consequences of their behaviour. Consequently, finding suitable candidates for a cohort study would not be easy. It might be possible to use the sad cases of infants who are known to the authorities to have been victims of sexual abuse. However, in these cases the extent of the trauma may be far greater than infants from otherwise caring homes who have occasionally been punished for their 'naughtiness', thus introducing variables not countenanced in the original comparison. But even if we ignore these kinds of practical problems and assume that two cohort groups of adequate size can be formed major problems facing this design still remain.

Cohort studies are not well-controlled, there is no randomization of subjects into two groups, and no employment of an experimental intervention. It is extremely unlikely that the two initial groups will be equivalent for all factors other than the test variable. But then we would not be able to rule out the possible influence of extraneous factors. This, in turn, would make an inference about the influence of the independent variable (i.e. sexual repression due to punishment) on the dependent one (occurrence of symptoms of obsessional neurosis) of questionable value. The potentially confounding influence of extraneous factors (e.g. maturation and external influences) would constitute a considerable risk, given the considerable time which elapses during a cohort study. I accept that the attempt to form two groups which differ 'only' with respect to the independent variable is a praiseworthy

undertaking and is to be welcomed. However, given the very high standards for the demonstration of causal relevance that Grünbaum has argued for, the stakes have been raised and it is now incumbent upon those who advocate the use of cohort studies to provide a defence against the charge that such studies might not be able to provide sufficiently rigorous tests of the aetiological hypothesis at issue. In particular, we would need to be persuaded that the potential liabilities of pertinent cohort studies are significantly less than those of case-studies and, moreover, that the conclusion afforded by the cohort study is 'beyond reasonable doubt'.

(ii) Randomized Clinical Trials (RCTs).

Even though, in principle, clinical trials have greater capability for reliably discerning a causal influence than cohort studies, there are obvious ethical prohibitions on carrying out even the early stages of a trial for the hypothesis at issue. Assume, just for argument's sake, that we had a sample of infants known to harbour infantile sexual inclinations (such as the tendency to fondle their genitals), none of whom had yet been reprimanded for their behaviour. It would be utterly unacceptable on moral grounds randomly to divide this sample into two groups and to attempt to enforce repression in one group by punishing them for their behaviour. For all practical purposes, therefore, in a society which observed even minimal standards of ethical conduct the experiment is unperformable. This is even though, in principle, and on epistemic grounds alone, it might have the capacity genuinely to discriminate some differences between the control and experimental groups following the intervention.

There is also the further problem that the dependent variable we are interested in (i.e. symptoms of obsessional neurosis) may not appear until very much later (e.g. in adolescence or adulthood), if it appears at all. Suppose, just for argument's sake, that the ethical prohibition could be lifted and that the experiment could be performed. Then, if systematic differences appeared between the two groups, say, 10 to 20 years after the intervention, could we be sure that these were due to the experimental intervention and not due to the influence of other variables operating in the interim? The potential threats to internal validity termed 'history' and 'maturation' by Campbell and Stanley ([1963] 5, 13-14) need to be considered. By 'history' is meant specific environmental influences operating between the time

of the experimental intervention and the time of the outcome observation. By 'maturation' is meant factors which are due to internal changes in the subjects which become manifested with the passage of time (e.g. emergence of latent personality traits). The standard RCT format (see especially designs 4 and 6 in Campbell and Stanley, op. cit.), if carried out under sufficiently ideal conditions should, in principle, overcome those difficulties. This is because, according to theory, the initial randomization should apportion to each group approximately equivalent numbers of maturation traits, and it is expected that history factors will not significantly affect either group more than the other. However, whether or not this theoretical ideal will be satisfied in practice is a different matter. An important factor is the size of the groups (see section 4.3(a)). If the groups are small, then history factors operating in the interim might substantially influence the result. Similarly for maturation. Even if the groups are not small the RCT design does not offer any way of checking whether history or maturation factors have biased the result in the long intervening period: faith has to be placed in the assumption that the initial randomization will suffice to exclude such a possibility.

Overall, Grünbaum is far too uncritical about the ability of prospective, controlled group-comparison methods to test the type of aetiological hypothesis at issue (see e.g. Grünbaum [1984] 258-259). Indeed, at times he speaks as if those methods are simply waiting to be applied and will provide an effective test on call (ibid.). But, as we have seen, this is unlikely to be the case. If the study cannot be conducted for ethical reasons then there can be no actual epistemic gain: the only test we need consider is a test *in practice*. But even if the test were allowed, considerable caution would need to be exercised in accepting a result given (at least) the threats from history and maturation. Beyond this, there is the problem of how the generic result obtained by the RCT method could answer the question about aetiology in the *individual* case that a psychotherapist might well be asking (see subsections 4.3(d) and 5.5 (B)).

CHAPTER 5

DIFFICULTIES IN EFFECTIVELY *APPLYING* A PRINCIPLE OF CAUSAL RELEVANCE

5.1 INTRODUCTION

The aim of this chapter is to examine further some of the problems involved with testing causal psychotherapeutic hypotheses. The specific focus will be:

(a) To provide a critical examination of a model of causal relevance espoused by Grünbaum, insofar as that model can be *applied* in practice for testing purposes; and

(b) To undertake the above specifically with regard to the testing of a certain category of unexceptional singular category II causal hypotheses with an intentional structure.

Hypotheses pertaining to single individuals or single occurrences which have an intentional structure have been chosen because they are common fare for psychotherapists. We need to consider what special problems arise for the testing of such hypotheses and, in particular, whether the model of causal relevance advocated by Grünbaum can be used effectively in practice to test them. I shall proceed as follows:

Firstly, in section 5.2, I shall introduce the model of causal relevance used by Grünbaum (I shall refer to it as the 'Standard Model of Testing for Causal Relevance' – or SMCR).

Secondly, in section 5.3, I shall present some of the case-study examples I have in mind. I shall draw attention to significant features of these examples, and shall identify the causal hypothesis which is relevant in each case.

In sections 5.4, 5.5 and 5.6 I shall examine whether the SMCR, when interpreted and applied in practice, can be used effectively to test the causal hypotheses of interest. These three sections will cover, respectively, the informal application of the SMCR, and what I shall refer to as the 'synchronic' and 'diachronic' experimental applications of it. My general conclusion will be that whether the application is informal or experimental, the problems of being able effectively to test the relevant causal hypotheses in practice by strategies closely based on the SMCR are substantial. Indeed, they are so substantial as to cast into

doubt whether the rigorous testing of those hypotheses by that formula is possible in practice unless, conceivably, effective designs other than those considered here could be found. Finally, in section 5.7, I shall present a discussion of some of the issues arising out of the preceding analysis.

5.2 GRÜNBAUM'S STANDARD MODEL OF TESTING FOR CAUSAL RELEVANCE (SMCR)

At several places in his writings on psychoanalysis Grünbaum proposes a model or schema for discerning causal relevance. For example, he says:

> "...I have repeatedly mentioned the following *necessary condition* for the *causal relevance* of an attribute or factor X to the occurrence of a property Y in a reference class C in which there are instances of X: X divides the class C into two subclasses, X's and non-X's, such that the respective probabilities of Y in these two subclasses are *different*...."
>
> (Grünbaum [1993] 163; emphases in original; reference omitted. Essentially the same model is reiterated in Grünbaum [1990] 570-571, [1994](a) 54 and [1994](b) 156-157.)

For brevity let us call this schema the 'standard model of discerning (or testing) causal relevance' (SMCR).

In order to avoid possible misunderstanding of Grünbaum's statement the following preliminary distinction will be helpful:

> (1) If factor X divides the class C into two subclasses, X's and non-X's, and if the probabilities of Y in these two subclasses are different, then X is (actually) causally relevant for Y.

> (2) If factor X divides the class C into two subclasses, X's and non-X's, and if the probabilities of Y in these two subclasses are different, then the necessary condition for X being causally relevant for Y is satisfied (but we still have not provided conditions *sufficient* for X to be causally relevant for Y).

Grünbaum cannot have meant '(1)', since a partitioning by X of C into subclasses differing in their probabilities of Y would not be sufficient for X's causal efficacy with regard to Y. To illustrate this, consider the following example. Suppose that the property of having long hair (X) - where 'long' can be defined as, say, over ten inches - divides a population of humans (C) into those with long hair (X's) and those who do not have long hair (non-X's). Suppose also that the probability of individuals with blue eyes (Y) in the first subclass of this population is

different from that in the second subclass. It obviously does not follow that having long hair in this population is causally relevant for having blue eyes! Consequently, I think Grünbaum must have meant something along the lines of (2). Thus construed, the SMCR provides a necessary condition (in the sense of a conceptual precondition) for the causal relevance of a factor X to a property Y, but does not inform us whether, in a given case, X is or is not in fact causally relevant for Y.

By itself, the SMCR is just an abstract schema – i.e. a way of conceptually explicating what it means for one factor to be causally relevant for another. But what ultimately counts is how it can be applied in practice and, when applied, how effectively it can be used to test the target hypothesis. The SMCR may not even be the only way in which the concept of causal relevance can be explicated. That is, there may be other ways in which a meaningful concept of causal relevance could be articulated which is not equivalent to (or reducible to) the SMCR. (We shall not, however, explore this latter possibility in this thesis.) The SMCR is certainly sufficiently important and wide-ranging to deserve scrutiny, and I shall proceed on this basis.

A central theme of this chapter is that there is a major difference between, on the one hand, providing an explication of the concept of causal relevance (which is what Grünbaum does in his formula) and, on the other hand, being able to apply that concept to test for causal relevance effectively in practice. Even if Grünbaum is correct in believing that he has provided a general formulation of the necessary condition for the causal relevance of one factor for another, it does not follow that he has provided an effective and practically applicable way of testing for causal relevance in the domains under consideration. The provision of an abstract conceptual formulation of causal relevance (even if that account is adequate) does not solve the problems associated with ascertaining whether one factor is in fact causally relevant for another. The latter question, and the practical problems associated with it are, however, methodologically crucial.

The SMCR does not assist us in being able to provide a test for causal relevance beyond being an essentially *a priori* schema that needs to be interpreted and applied. When it is interpreted and applied there is no guarantee that an effective test of the target hypothesis will be effected through its utilisation. As will be seen in sections 5.4, 5.5 and 5.6, there are many potential impediments to its serving as a basis for an effective test of the hypotheses in question.

5.3 THE EXAMPLES

The type of hypotheses I shall be interested in involves the possible causal relevance of some external or 'worldly' circumstance for a cognitive-emotional state with an intentional (or representational) structure. In each case the external circumstance that will be considered as the possible causally relevant factor (or independent variable) will be the 'worldly counterpart' (if it exists) of the intentional content of the cognitive-emotional state (the dependent variable). To illustrate and explain this further consider the following example:

> 'Jane is distressed about being treated with aloofness and neglect by her father'

If we analyse this, it can be seen to have the structure: 'A_1 is Y_1-about-X_1'; where A_1 = Jane; Y_1 = Jane's emotional state of distress; X_1 = being treated with aloofness and neglect by her father. It is to be emphasised that X_1 refers to what Jane is distressed about, in the sense of the intentional (or representational) content of her distress. It would be premature at this stage to conclude that Jane's father is treating her (or has treated her) with aloofness and neglect. A well-recognised feature of intentional content is that it may not refer to anything real (see e.g. Føllesdal [1982]; Searle [1983] 4); and establishing whether Jane's father does treat her in that way (or has done so) is an independent matter. Let us now use the symbol X_1* to refer to Jane's father's treatment of her with aloofness and neglect *if that is actual*. That is, X_1* refers to the 'worldly counterpart' of the intentional content, X_1, just in case there is a circumstance in the real world corresponding to it. We are now in a position to formulate the causal hypothesis in which we are interested:

> α_1: 'Jane's father's treatment of her with aloofness and neglect (if actual) is positively causally relevant for her feeling distressed about being treated with aloofness and neglect by her father.'

α_1 therefore has the structure: X_1* (if actual) is positively causally relevant for Jane's Y_1-about-X_1. α_1 is the hypothesis we are interested in, and our testing problem concerns whether the SMCR can be used to test it. α_1 is, of course, just one example from a class of hypotheses with a similar structure (let us call this class $\hat{\alpha}$, where $\hat{\alpha} = \{\alpha_1, \alpha_2, \alpha_3 \ldots \alpha_n\}$). Other examples of particular cognitive-emotional states could be considered:

'Tom feels guilty about socially disowning his father'
(see chapter 3)

'Nick is angry about his wife's infidelity'
(We could even suppose that it is uncertain that Nick's wife is being unfaithful. Let us assume that there is no unequivocal evidence for this, and that people who know Nick well believe his suspicions to be without foundation, and that he has become somewhat unbalanced and is suffering from a so-called 'Othello Complex'.)

For each of these, the corresponding causal hypothesis could be formulated:

α_2: 'Tom's actions of socially disowning his father (if actual) are positively causally relevant for his feelings of guilt about socially disowning his father'

α_3: 'Nick's wife's infidelity (if actual) is positively causally relevant for his feelings of anger about his wife's infidelity'

As with the first example involving Jane, it is essential in these (and other similar) examples not to assume that the respective intentional objects (i.e. representational contents) of the cognitive-emotional states necessarily correspond to any real worldly state-of-affairs. Independent enquiry – and relevant supporting or disconfirming evidence – is required in each case to ascertain whether or not the 'worldly counterpart' obtains. To keep our testing problem comparatively simple I

shall confine the main discussion only to cases in which the worldly counterpart, X_i*, is actual. However, from a psychiatric point of view some of the most interesting examples are when X_i corresponds to nothing real (i.e. there is no corresponding X_i*). This, of course, includes examples of delusional belief in which it is of considerable theoretical interest to consider what role (especially causal role) might be played by the delusional content in the agent's overall pathology. The example of Nick illustrates this as does, with even greater force, the further example of Albert. Albert, let us suppose, is suffering hypochondriacal anxiety about having a malignant brain tumour when he is in fact completely free of cancer:

> 'Albert is anxious about having (and not just about being potentially susceptible to having) a malignant brain tumour'

In this case, the presence of a tumour in Albert's brain (i.e. what would have been represented by X_4*) clearly cannot be causally relevant for his anxiety about having a brain tumour, since he does not have a tumour (i.e. X_4* does not exist). Albert's anxiety (i.e. Y_4) is nevertheless real; and so is the subjectively experienced 'meaningful' content of what he is anxious about (i.e. X_4 = his having a brain tumour).

A further point worthy of note is that the 'worldly counterpart', X_i*, does not have to be restricted to states-of-affairs external to (in the sense of being physically separate from) the agent. Instead, it could include the agent's own actions, behaviour or physiology. For example, in α_2 the worldly counterpart consists primarily of Tom's own actions of, for example, avoiding being seen in public with his father.

To summarise, the general form of the hypotheses that we shall be interested in testing is as follows: X_i* (an external or 'worldly' circumstance) is causally relevant for Y_i-about-X_i (i.e. a cognitive-emotional state, Y_i, with representational content X_i). We shall limit the discussion to cases in which X_i* exists – i.e. X_i corresponds to something real.

5.4 CAN THE INFORMAL APPLICATION OF THE SMCR TEST α_1 EFFECTIVELY?

We shall use α_1 as a specimen example of the class of hypotheses $\hat{\alpha}$ (containing hypotheses α_1, α_2, α_3...etc.).

It might be felt that a straightforward way in which the SMCR could be applied to test α_1 is informally, as follows:

Take careful note of Jane's psychology (as understandable in FP terms) following interactions with her father. A record will need to be kept of those instances in which Jane's father treats her with aloofness and neglect and those instances in which he does not. Careful note will also need to be taken of Jane's emotional state following the interactions. Are the occasions of her experiencing distress at being treated with aloofness and neglect by her father of a different frequency (or different intensity) when he does treat her in that way as compared to when he does not? If there is a difference in frequency or intensity between the two cases we would infer that the aloof and neglectful behaviour is making a statistical difference to Jane's distress. This data could then be used to reach a judgement about causal influence.

A distinction needs to be made between this informal but consciously applied use of the SMCR to test α_1 and the use of FP-reasoning (e.g. in a counselling session) to do so. In the former, a known strategy based on the structural features of the SMCR is consciously applied. In the latter the reasoning is more fluid and uncontrolled and, as I have suggested elsewhere (see pages 38, 171, 195-196 and 228-229), we in general possess no precise knowledge of what inferential strategies are used, or how FP-reasoning operates.

Ordinarily, we would expect that being treated with aloofness and neglect is *positively* causally relevant for feeling distress at being treated in that way. This is what naïve folk psychology seems to teach us. However, even when employing the SMCR informally, as above, we should be careful not to prejudge the issue.

That Grünbaum believes that an informal use of the SMCR is sometimes warranted is evidenced by an example (Grünbaum [1993] 164) in which he uses it in that way to infer the emotional consequences of being insulted (I

discuss this example at some length in section 7.3). However, there are significant problems which face the informal application of the SMCR (this applies generally, as well as specifically for the purpose of testing α_1).

Firstly, as mentioned earlier (see pages 128, 140-141 and 143), Grünbaum has substantially raised the stakes with regard to the standard of demonstration for causal relevance claims. This applies now as much to conclusions reached on the basis of an informal use of the SMCR as to an experimental (or quasi-experimental) application of it.

Secondly, in an informal application of the SMCR the numbers of occasions on which the independent variable is or is not applied is a matter of happenstance. A sufficiently large number of instances of each type (i.e. presence vs. absence of the independent variable) would be required for the inference. For example, if Jane's father treated her with aloofness and neglect on *every* occasion on which they encountered one another, there would be no contrast group with which to make the inference.

Thirdly, when the SMCR is used informally the inference made by means of it is potentially vitiated by certain data-contaminating effects. For example, we cannot be sure that when, on one encounter, Jane's father treats Jane with aloofness and neglect, the effect of his behaviour on her psychology will not persist and influence her later responses to him. Let us call this a 'persistence of response' effect. To illustrate this, suppose that on an initial occasion Jane's father treats her with aloofness and neglect and that she subsequently experiences distress at having been treated in that way by him. Suppose also on their next encounter he behaves towards her in the same way and that she again experiences distress of the same type straight afterwards. How can we be sure that her distress at being treated with aloofness and neglect on this second encounter is (causally) due to the second exposure to aloof and neglectful behaviour, as opposed to being an enduring psychological effect due to the first exposure? It is, after all, possible that having been once (or several times) treated with aloofness and neglect by her father that, on a subsequent encounter with him, Jane responds with distress of the specified kind not because of the new exposure, but because of a lasting influence from a previous exposure which is triggered by some non-specific aspect of the fresh encounter (e.g. by simply being in her father's presence; or memory of a previous occasion on

which he treated her in that way). This is an alternative hypothesis which deserves serious consideration.

The above point is of methodological importance because of the way in which the SMCR is meant to work when applied informally. In the informal application of the SMCR as described, a numerical count is made of what the subject's response was in cases of repeated exposure to the independent variable, and this is compared with her response on multiple occasions when she was not exposed to the independent variable. The respective relative frequencies which are generated form the statistical basis for making a judgement about causal efficacy. The method therefore makes crucial use of the numbers involved, and there is a tacit assumption that however the subject responds to a given exposure to the independent variable, this was not affected by previous exposure(s), and will not influence future responses. That is, there is an assumption that applications of the independent variable and any subsequent change in the dependent variable can be treated as discrete pairs of occurrences which are *causally independent* of other (earlier or later) pairs of that type. This assumption is required if the method of relying on making a numerical count of the dependent variable outcome under, respectively, exposure and non-exposure to the independent variable is to be valid. But it is just this assumption of causal independence between pairs of occurrences that will be violated if the 'persistence of response' effect is real in a given case. This, in turn, could lead to a miscount, resulting in an error in the conclusion drawn on the basis of the numbers used.

In the problem that has just been discussed the dependent variable was present when the independent variable was present, but the former's presence was due to a previous exposure to the independent variable. In the fourth and final problem I shall discuss the dependent variable is present when the independent variable is, but the former's presence is due to a factor altogether unrelated to the independent variable (in my example, endogenous neurotic anxiety). This, again, could lead to a miscount.

Suppose we observe a series of encounters between Jane and her father. In some of these he treats her with aloofness and neglect and she experiences distress at being so treated almost immediately afterwards. With no contrary background information to go on, we would ordinarily class these cases – when

using the SMCR informally – as evidence for a relation of positive causal relevance of the independent variable for the dependent one.

However, suppose that on several encounters Jane experiences the specified distress even when her father does not treat her with aloofness and neglect, or that she even experiences it when she is on her own. One possible explanation for this would be a 'persistence of response' effect of the type mentioned in the previous example (this would, of course, only be plausible if Jane had been exposed to the independent variable on at least one prior occasion). This, however, is not the only explanation. Suppose that Jane is neurotically anxious about being uncared for, such that she is susceptible to experiencing the specified type of distress even in the absence of a corresponding environmental stimulus (i.e. her distress has an endogenous or intrapsychic source). If this is the case then, once again, there is a risk of misattributing the presence of the dependent variable to the presence of the independent one. Even though, let us suppose, Jane's father treats her with aloofness and neglect on a particular occasion and she experiences distress at this soon afterwards, we cannot be sure that her experience of that distress was causally attributable to her father's behaviour rather than to her neurotic disposition. An informal usage of the SMCR, in which we are simply making a count of the presence/absence of the two variables (or their intensities) during encounters between Jane and her father does not enable us to distinguish between these alternatives. What is needed is some way of screening-off cases when the dependent variable is present because of neurotic anxiety rather than because of exposure to aloof and neglectful behaviour. There may be ways of doing this. For example, it might be possible to make a separate assessment for neuroticism or the harbouring of specific endogenous anxieties and to compensate for these effects. However, because making allowance for such factors would amount to introducing measures of control, this would tend to make the study a quasi-experimental one rather than its being an informal application of the SMCR.

In conclusion, whereas it might initially have seemed that the SMCR could be used simply and effectively in an informal way to test α_1, this is unlikely to be the case.

It might be felt that the effective testing of α_1 could, instead, be achieved by an experimental application of the SMCR. In the next two sections I

shall examine this possibility. First, I shall apply the SMCR in what is essentially a basic group-comparison format. Because the groups are compared simultaneously I have called this a *synchronic* application of the SMCR. The second experimental application is modelled essentially on a standard time-series (or AB) format. Because the contrasted sets of data are temporally ordered (within a single individual) I have called this a *diachronic* application of the SMCR.

5.5 CAN THE SYNCHRONIC EXPERIMENTAL APPLICATION OF THE SMCR TEST α_1 EFFECTIVELY?

Experimental applications of the SMCR would require subjecting test individuals to a negative intervention and, for this reason alone would be ethically impermissible. However, if we were not allowed to think through the steps involved in carrying out the experiments we would not be able to analyse the methodological problems involved in attempting to test α_1 experimentally. For this reason we shall proceed hypothetically, as if we could perform the experiments, our aim being to expose the epistemological difficulties. Amongst the difficulties encountered are ones of an essentially practical nature. But these, as I have suggested before, should also count as epistemological if they interfere in a fundamental way with the information that is capable of being extracted from the potential test situation. Ultimately, testing has to be testing in practice or it amounts to nothing. If a hypothesis cannot be tested in practice the test of that hypothesis fails epistemically.

In order to apply the SMCR experimentally and synchronically we would need to form a group of n suitable women whose fathers treated them with aloofness and neglect and subsequently observe how many did/did not feel distress at being treated in that way. Let us suppose that there were q out of n women who did feel such distress. Also, we would simultaneously need to run a control group of m suitable women whose fathers did not treat them in that way and note how many of these control group subjects experienced distress of that type (suppose there were r). We would then need to compare the experimental and control group frequencies: respectively, q/n and r/m. I shall completely gloss over the (very substantial) problem of how causal conclusions can be derived from statistical data[38]. I shall assume that q/n is much greater than r/m, and that this would allow us to infer that, in the sample used, aloof and neglectful behaviour by fathers is positively causally relevant for daughters experiencing distress at being treated in that way.

Even if the experiment were ethically permissible there would be certain preliminary difficulties involved in being able to obtain reliable values for q/n and r/m. I shall deal with these in subsection (A), below (the diachronic experimental application also faces some of these difficulties – see 5.6). The main

focus of this section will, however, concern the problem of whether the synchronic application can be relevant and effective for testing a hypothesis about a single individual (subsection (B)).

(A) Preliminary Difficulties

The preliminary difficulties are:

(i) Could suitable samples be obtained, and could the experiment be organised in practice?

Since our aim is to reach a conclusion about, specifically, Jane and her father, the closer the match between the relevant characteristics of the father-daughter pairs used in the experiment and those of Jane and her father, the better. Insofar as those attributes (e.g. age, personality-type, nature of father-daughter bond etc.) increasingly differed from those of Jane and her father, any inference as to the likely causal influence of Jane's father's behaviour on her emotional state based on the experimental data would increasingly tend to err. Being able to find a sufficiently large sample of suitable father-daughter pairs would not be easy (see 4.3(a)), but it might not be impossible. As for the practical organisation and implementation of the experiment, various difficulties present themselves, but these might not be insuperable. The following plan is suggested:

Fathers and daughters having characteristics similar to Jane and her father (for factors likely to affect the outcome) would be used in the study. It would be highly preferable for the study to be carried out without the daughters' knowledge in order to reduce the risk of 'reactive' effects (this will be explained in point (iii), below). (This latter requirement would, of course, be ethically dubious, again raising the question of whether the experimental is permissible.) Daughters would be assigned at random to the treatment and control groups, and those fathers in the treatment group would be expected deliberately to behave in an aloof and neglectful way towards their daughters for a certain period of time. Control group fathers would behave as per normal. In order to keep background factors as constant as possible the experimental intervention could be carried out at home at an appointed time, and under domestic circumstances that were as little changed as possible from the usual run of things. Following the intervention an assessment

would have to be made for the presence/absence of the dependent variable (or its intensity), and a similar evaluation made for the control group.

(ii) Could the independent variable be implemented in practice, and in a way that was realistic?

The experimental intervention would not be an easy task for fathers to perform, even if there was no moral prohibition on the action. Fathers might not be able sincerely to treat their daughters with aloofness and neglect if they had not already been accustomed to doing so. And that fathers did not have a prior history of treating their daughters with aloofness and neglect would be a prerequisite for the entry of a father-daughter pair into the experiment (otherwise there would be a high risk that the daughter would be 'contaminated' for the effect we were attempting to discern). In order for fathers to behave in the required way they might well have to act in a manner that was contrary to their own strong positive emotional ties to their daughters. This is significant because the paternal behaviour exhibited as the experimental intervention would have to be realistic – i.e. behaviourally very similar to such behaviour as it occurs naturally (and, in the particular case at hand, as displayed by Jane's father). If the intervention behaviour deviated significantly from what occurred naturally it would jeopardise any conclusion reached about the real-life case that was based on the experimental data. However, for fathers who had not habitually treated their daughters with aloofness and neglect to do so in a manner that was realistic, and on call, would require uncommon skill on their part. A second point, related to the issue of reactivity (to be dealt with under point (iii), below), is that any performance short of a fully convincing one might alert daughters that something untoward was afoot. Kazdin provides an example of a study in which he says it was obvious that "something was not quite right and that [it was obvious to onlookers that] an experiment was in progress" (Kazdin [1980] 267; see also op. cit. 239).

A further problem concerns the standardisation of the experimental intervention. Our original question was whether Jane's father's aloof and neglectful behaviour was causally relevant for Jane's distress (at being treated in that way by him). The relevant behaviour on the father's part would have consisted of specific events in specific settings (e.g. Jane's father failed to give Jane attention or comfort when she suffered a minor injury). Since we are ultimately interested in an

individualised (or singular) hypothesis about Jane and her father, if we are even to attempt to make an inference about that case from a study involving other father-daughter pairs, it is essential that the intervention administered in these other cases is as similar as possible to that which occurred in the original pair. In other words, we need to reproduce for the subjects participating in the experiment, at least approximately, the kinds of behaviour that constituted aloof and neglectful behaviour in the original. This might be difficult or even impossible, and constitutes a serious constraint on the performability of the experiment. If, for instance, during the period in the experiment allotted to exposing the daughter to aloof and neglectful behaviour she did not suffer a minor injury, then the opportunity for the father to fail to give her attention or comfort - see above - would not arise. However, we do not need to be absolutely strict about this requirement. Being so would render the experiment unperformable, since an exact reproduction of the original circumstances would in any case be impossible. It is hoped that what constitutes aloof and neglectful behaviour can be sufficiently clearly defined and yet has sufficient flexibility to permit of its experimental implementation. For example, a general emotional detachment ('coldness') on the father's part might be sufficiently general to be experimentally implementable, and yet sufficiently akin to Jane's father's behaviour to be relevant for testing its effect on Jane. Overall, I think the problem of standardisation is a substantial one, but not one that we should see as inevitably rendering a relevant experiment unperformable.

(iii) Would test-subjects react to the assessment procedures and experimental intervention in a way that would vitiate the conclusion drawn?

In an experiment as ideally performed the method of assessing the dependent variable should, so to speak, be invisible and should not obtrude upon the effect which it is the aim of the experiment to detect. Similarly, the process of administering the independent variable should not have any secondary effect on the test-subjects which interferes with any potential effect which that variable, *per se*, has. 'Reactive' effects are ones produced by test-subjects' responses to, specifically, the method of assessment, or as an unwanted concomitant of the application of the experimental intervention (see e.g. Kazdin [1980] 45, 256-257, 260-261).

Kazdin says:

> "Reactivity of assessment refers to the fact that subjects may be aware that their personality or behaviour is being assessed and that their responses are altered as a result....
>
> ...A host of other influences, extending beyond assessment, are related to the reactivity of the experimental situation itself....
>
> ...Whether reactivity pertains to assessment or to the experimental arrangement, it raises the possibility that subjects perform in a particular way unrepresentative of their performance under nonreactive conditions."
>
> (Kazdin op. cit. 257, 260-261)

In the present (hypothetical) study there is a risk of reactive effects confounding the conclusion drawn.

Earlier (in 5.5(A)(i)), it was suggested that it would be highly preferable for daughters not to be informed about their involvement in the experiment. This is because if they did know, it would almost certainly have a major influence on their response. Typically, it might be expected that such prior awareness would drastically lessen the impact of the intervention when it occurred. Daughters might not take the intervention seriously, or might regard it as a temporary 'performance' from which to remain emotionally detached. Alternatively, they might respond powerfully (e.g. with anger or feelings of betrayal) not to the intervention itself but, for example, to the fact that their fathers had been willing to participate in such a callous experiment. Some might even respond with hilarity at the prospect that their fathers would soon have to put on a sustained display of aloof and neglectful behaviour.

Reactivity to assessment of the dependent variable would, potentially, be a serious problem. This would especially be the case if a pre-test assessment were undertaken, since that would risk inadvertently priming the test-subject for the intervention to follow. A pre-test (or pretest) assessment is an observational evaluation (or measurement) of the dependent variable prior to the experimental intervention. Typically, the pretest will be carried out after the randomization, but it can be carried out prior to the randomization (Kazdin [1980] 131). There are ways of attempting to minimise this difficulty: for example, by attempting to evaluate the dependent variable under the pretext of carrying out an interview on some totally unrelated and innocuous subject. Alternatively, an evaluation of the influence of the pretest on the outcome could be made using a Solomon Four-Group Design (see e.g. Campbell and Stanley [1963] 24-25) which, in addition to having the experimental and control groups with pretest evaluation, has further experimental and control groups without pretests. Some commentators even call into question the

methodological value of pretests. For example, Campbell and Stanley (op. cit. 25) argue that in genuine experimental studies the pretest serves only the psychological function of reassuring researchers that the compared groups are equivalent, whereas this end is in fact methodologically served by the randomization. Others, however, value the pretest not only as a way of ensuring that the groups are equivalent, but as a means of gathering more general data about subjects so that they can be matched prior to randomization (e.g. Kazdin [1980] 131-132). Nobody, however, would deny that administering a pretest risks some kind of reaction (Kazdin ibid.). In the present (hypothetical) study some kind of preliminary examination would be essential (even if this was only to ensure that, prior to randomization, the father-daughter pairs participating in the study did not differ markedly in their characteristics from Jane and her father). Thus, even if a pretest assessment for, specifically, the dependent variable were waived there would still be some risk of a reactive effect prior to the experimental intervention.

Reactivity effects at post-test might also be significant. A post-test (or posttest) assessment is an observational evaluation (or measurement) of the dependent variable following the experimental intervention. For example, it might not be possible to elicit from the daughter a completely open and accurate statement of her feelings under conditions in which the assessor (presumably, a trained observer) had not been taken into confidence. The latter might not be possible if only one or a few posttest interviews were scheduled.

The risk of a reactive response to the experimental intervention is significant and relates to the issue of the realism with which that intervention is administered, discussed in point (ii), above. It is, I think, fair to assume that psychological responses to genuine or spontaneous behaviour in others can be very different from responses to behaviour which is perceived as not genuine. If the behaviour is not spontaneous but enacted, but the enactment is very realistic, the subject may never realise that it was a performance rather than genuine. On the other hand, subtle if unintended cues in non-genuine behaviour can betray its inauthenticity. I also assume that humans are generally very sensitive to nuances which distinguish genuine from faked behaviour, perhaps especially in persons whom they know closely (for example, parents). In the present experiment, anything short of a near perfect enactment of the requisite aloof and neglectful behaviour by the father could result in a response which differed from that which would have

resulted had the behaviour been spontaneous. The daughter might well respond to unintended concomitants of the enactment (e.g. hesitancy or a guilt-reaction of the father) rather than to behaviour perceived by her as genuinely aloof and neglectful. This could easily vitiate any general conclusion drawn from the study about the influence of aloof and neglectful behaviour, as well as the extrapolation of that conclusion to other cases.

(B) Testing the Individual Case

The logic of the synchronic experimental application is based upon a contrast between two groups (of at least one member per group). However, the hypothesis, α_1, which we wish to test is about a single individual, Jane. If it were required that α_1 had to be tested directly by the synchronic application of the SMCR, then the fact that we have only one subject in all would stop the test dead in its tracks.

The synchronic study deals with a *generic* causal claim. This applies either to the specific sample used in the study or, if it is theoretically justified to regard that sample as being representatively drawn from an ideal statistical population, then it will be possible to extrapolate the result to that population. In either case, what holds for a sample/population does not necessarily hold for a particular individual (even if the individual happens to be a member of the sample used). The situation is roughly analogous to, for example, testing whether 'Ibuprofen' is causally relevant for relieving headaches (a generic hypothesis) when what we want to test is whether 'Ibuprofen' is causally relevant for relieving (say) Michael Dash's headaches (an individual or singular hypothesis). (I have used this example because as a matter of fact 'Ibuprofen' seems to have little or no efficacy for me, whereas it is has been proven to be an effective headache-reliever on the basis of generic studies.) Since, for practical reasons, the synchronic experiment cannot be applied directly to test α_1, perhaps the best we can hope for is some way of relevantly applying the experimental result to reach a conclusion about whether, for Jane, her father's behaviour is or is not causally relevant for her distress. On the generous assumption that the problems mentioned under '(A)', above, could be circumvented, let us suppose that the experiment was carried out and that it provided

us with the following empirical law, [T] (N.B. [T] is based on the assumed result that q/n >> r/m – see page 158):

> [T]: 'Daughters who have been treated with aloofness and neglect by their fathers are very much more likely to experience distress at being treated in that way than daughters who have not been so treated'

Once again glossing over the difficulties of deriving causes from frequencies, let us assume that [T] entitles us to a conclusion, [U], about causal relevance:

> [U]: 'Fathers' aloof and neglectful behaviour towards their daughters is (positively) causally relevant for their daughters experiencing distress at being treated in that way'

[U] is a generic causal conclusion. Can it be used to reach a reliable conclusion about whether (or not) Jane's father's treatment of her with aloofness and neglect is causally relevant for the distress she experiences at being treated in that way by him? This is the central problem we are faced with, given we are supplied with a result from a group-comparison experiment, but want to know about causal influence in an individual case.

We would expect that, *ceteris paribus*, the more alike in their attributes that the subjects participating in the experiment were to Jane, the more reliably would we be able to draw a conclusion about her on the basis of the experimental result. This is generally true. If the participants were exact replicas of Jane and if the experimental intervention accurately replicated Jane's father's behaviour, then the inference to the particular case of Jane would be very reliable. However, being able to obtain subjects who closely matched Jane for factors potentially influencing the outcome would in practice be very difficult. Any inference from a group result to an individual case in which (i) the group members differed from the individual, and in which (ii) the inference was not made deductively would lead to some inductive uncertainty about the conclusion drawn. Since it would be virtually impossible to satisfy either of those conditions (i.e. exact

replication, or deducibility), it is almost inevitable that the drawing of a conclusion about an individual from a generic premise will involve an unspecified or unspecifiable *judgmental* component, and will risk error.

Whereas the risk of error is ever present in inductive reasoning, the above judgmental component goes against the spirit of eliminating subjective bias which is an important reason for relying on experimentation in the first place. Let us not forget that experiments are often promoted under an image of objectivity (see 4.3(f)). The performance of all experiments will involve interpretation and judgement at certain points, but there is surely a difference between this and, as in the present case, a conclusion that is primarily founded on a personal judgement (or a set of such judgements). The difference is due to the fact that the synchronic experimental format cannot be applied directly to test α_1, so that any conclusion drawn about that hypothesis which uses the experimental result will be of the nature of such a judgement. This needs to be contrasted with experiments which test hypotheses directly and which, at their best, are relatively objective in spite of an inevitable background role for judgement and interpretation. For example, a well-conducted trial to test the generic hypothesis that, say, a certain fertiliser increases the yield in tomato plants would be comparatively objective; as would Newton's classic prism experiment in which white light was interpreted as being separated into its constituent colours.

If we are to be fair in evaluating whether or not an experimental application of the SMCR can effectively test α_1, we must do so under the assumption that the conclusion drawn will inherit that degree of objectivity which is provided by the experimental set-up. In the present case, in which the generic experimental conclusion is merely used as a premise in a 'clinical judgement' about α_1, this condition will not be met. This is a further reason for concluding that the synchronic experimental application of the SMCR fails effectively to test α_1.

One type of response to what has been said so far would be to say that insofar as we expect group-comparison studies to be able to test individualised (i.e. singular) hypotheses we misconceive them. According to this view, group-comparison studies are by their very nature limited to testing generic hypotheses about samples or populations, and if they are applied for testing hypotheses about single individuals they are *mis*applied. This view is not unreasonable but, by

stipulatively disqualifying their application to hypotheses about single individuals, the range of experimental formats which remain which could potentially be used to test such hypotheses is correspondingly narrowed. This certainly cannot help to strengthen the case of those who believe that α_1 can be effectively tested experimentally. Moreover, it does not do full justice to the fact that results from group-comparison studies sometimes are used to reach informed judgements about individual cases.

Even if it were misconceived to employ a group-comparison experiment to test an individualised hypothesis, it would not follow that it would be misconceived to use an experimental application of the SMCR to do so. This is because the SMCR can be interpreted and experimentally applied in a different way to the group-comparison (i.e. synchronic) format – i.e. in accordance with what I have called the diachronic application. I shall examine this in the next section. Whereas the diachronic format is specifically designed to operate on a single individual (thus obviating the above objection), we shall see that it is no better at providing an effective experimental test of α_1. Thus, our conclusion will be that neither experimental approach provides a truly effective test of α_1.

5.6 CAN THE DIACHRONIC EXPERIMENTAL APPLICATION OF THE SMCR TEST α_1 EFFECTIVELY?

The typical experimental situation I have in mind is as follows:

An individual is monitored for an initial period of time prior to time t_x with respect to the dependent variable. The presence/absence or intensity of the dependent variable during this period is used as a 'baseline' (i.e. control). The experimental intervention (in this case aloof and neglectful behaviour by the father) is introduced at time t_x. Further assessment of the dependent variable is made subsequent to this intervention. In this case the contrast classes of the SMCR can be thought of as the two sets of time-slices in the periods respectively before and after the intervention. That is, we are comparing the level of distress (at being treated with aloofness and neglect by the father) in the time slices before the intervention with those subsequent to it.

Because the basic comparison is made across a period of time within a single individual's lifespan I have referred to the model as 'diachronic' (it falls into the category of within-subject single-case designs – see section 4.2). The arrangement described above, in which there are only two contrasted phases (i.e. before and after the intervention) is, in fact, a weaker version of the most common strategy of this type, the ABAB design. In the ABAB design an initial monitoring phase (the first 'A') is followed by the experimental intervention, after which the first 'B' period commences. This is then followed by the removal of the experimental intervention, marking the commencement of the second 'A' period, and so on. The strength of the causal inference when using the ABAB design comes from predicting that when the experimental intervention is made, there will be a marked change in the level of the dependent variable, and that this will then *revert* to the baseline level when that intervention is removed or naturally loses its potency (Kazdin [1980] 173). It is also normally expected that a second experimental intervention will reinstate a second 'B' period. Whereas this design is effective for testing certain variables (e.g. a drug administered to reduce high blood pressure) it cannot be effectively used if the experimental intervention has a lasting effect, or risks having such an effect. In that case we cannot utilise the added inferential power

of the design which comes from potentially observing a reversal of the effect after the first 'B' phase. Unfortunately, in the present example just this limitation applies. Even if it were ethically permissible to carry out the experiment, there would be a high risk that a single intervention might have a lasting effect (see 5.4). It would be unrealistic to expect the subject's emotional responses to revert to 'baseline' having been exposed to a period of (atypical) aloof and neglectful behaviour by her father. Even so, the experiment could, in principle, be used in the weaker AB form of a single intervention following baseline measurement, and we shall proceed on this assumption

If we were to carry out the experiment we could not use Jane as the test-subject since she would be 'contaminated' for the effect which the experiment was aiming to measure (i.e. distress of the type in question). This leads to what I view as the most important problem of testing α_1 by this model. Namely, that to carry out an effective test we would need a test-subject who was sufficiently like Jane in all relevant characteristics so that the conclusion reached could genuinely be said to be about α_1, and not about some other, possibly similar (but in fact significantly different), hypothesis α_1'. The relevant characteristics are, of course, those that are potentially causally relevant for the outcome (e.g. personality type, emotional make-up, age, I.Q., prior exposure to emotionally relevant stimuli etc.). It should be remembered that we are addressing the question of whether α_1 can be tested effectively. That we can effectively test some other hypothesis, α_1' (or yet others, α_1'', α_1''' etc.) is, for the purpose of this question, beside the point. Ideally, we would like a clone of Jane, alike her in all internal characteristics, and having been subjected to identical environmental influences with the sole exception of not having been treated with aloof and neglectful behaviour by her father.

Indeed, the problem is even worse since unless the father in the experiment was identical to Jane's father (with the exception that he had not previously treated his daughter with aloofness and neglect), we could not be completely confident that the experimental intervention was the only (potential) causally relevant factor introduced into the situation which influenced the outcome. Perfectly controlled experiments are, of course, unattainable in practice. However, using experimental subjects who differed from the originals would risk giving an erroneous causal conclusion. What needs to be considered is whether the father-

daughter pair who were enlisted into the experiment would be sufficiently similar to the original pair to make the scale of any resulting error fall within acceptable bounds. Also, given the risk of such experimental error, we need to consider whether this is less or greater than the risk of an erroneous causal conclusion which might result from using counselling methods (e.g. if Jane and her father were interviewed in GPC sessions). As I have argued elsewhere (see pages 32-33 and 45-46), it should not be assumed that an experimental format will necessarily be able to test a given hypothesis more effectively or reliably than informal methods.

The practical implementation of the diachronic experiment could be carried out in a manner similar to that of an *individual* participant in the experimental group of the synchronic experiment (see 5.5). Problems similar to those we encountered for the synchronic format – i.e. regarding the implementation and realism of the independent variable, and reactivity in the posttest – would recur for this format. Since these have already been discussed (section 5.5(A)), there is no need to repeat them here. However, it is worth emphasising that in the present case, unlike the synchronic case, a pre-test assessment of the dependent variable would be *essential*. Moreover, this monitoring would need to be comparatively protracted in order to establish a baseline for future comparison. The risk of this producing a reactive response in the test subject would be considerable.

All in all, there are substantial hazards to testing α_1 effectively by the diachronic format.

5.7 DISCUSSION

What general conclusions can be drawn from the preceding analysis?

Firstly, although the SMCR is an important model for testing causal relevance it may not be the only one (see 5.2), or necessarily the most effective for a given causal hypothesis in a given setting. (I shall not explore these issues further.)

Secondly, although we may conjecture that some naturally occurring FP inferences are based on the SMCR, it should not be accepted as fact that they are without appropriate evidence. This relates to the point (see pages 38, 153, 195-196 and 228-229) that we currently do not have a precise understanding of how FP-reasoning (as a natural process) operates or what inferential strategies it employs.

Thirdly, we have seen that there are substantial problems in being able effectively to test α_1 through applications of the SMCR. The liabilities of the informal application (5.4) are not, perhaps, surprising. Of greater methodological significance are the problems encountered with the experimental applications (5.5 and 5.6). It cannot, of course, legitimately be claimed that there is no experimental application of the SMCR which could test α_1 effectively. Only two broad (albeit wide-ranging) approaches were considered. Furthermore, it is not possible to predict what ingenious applications future experimenters might devise.

Nevertheless, we should at the same time prepare ourselves for an altogether different scenario: namely, that there are in practice certain causal hypotheses in the psychological and social domains (and possibly outside them too) which are simply not amenable to being tested rigorously by experimental formats, or by the SMCR in particular. These hypotheses might include the class $\hat{\alpha}$, as well as others besides. It may just be a 'meta-fact' about the practical manipulability of the world that the range of causal hypotheses which are effectively testable by experiments, or by the SMCR, is circumscribed.

A fourth point is that if effective testing fails it should not automatically be assumed that the failure lies in a shortcoming of the formulation of the hypothesis, or that it was because the hypothesis was not suited for being tested by that method. As long as an hypothesis is empirically well-formed (e.g. is coherent, pertains to theoretically viable and empirically detectable constructs etc.) it

should normally be acceptable as a candidate for testing. Equally, if effective testing fails it should not be assumed that the failure arose because an inappropriate method was applied to test the hypothesis. The latter *might* have been the case, but then it might have been because the method was inductively incompetent rather than because of a misapplication. In chapter 6 I shall critically discuss one response to the apparent failure of the experimental applications of the SMCR effectively to test α_1: namely, the claim that such tests are essentially scientific, and that science is not interested in individualistic claims such as α_1.

So far we have considered problems of testing α_1 through applications of the SMCR, especially experimental applications. But how would the testing of α_1 fare if it was put in the hands of a psychotherapeutic counsellor?

Firstly, it does not of course follow from the limitations of experimental tests, or of tests based on the SMCR, that the informal testing of α_1 in GPC sessions will be any more effective. Also, it must be acknowledged straight away that the latter informal method does not meet the standards required for scientific validation (see pages 99 and 232-234, and section 2.2). Nevertheless, let us attempt to evaluate how well it would do.

I shall assume that the counsellor would give Jane ample opportunity in the sessions to express and articulate her feelings of distress. I shall also assume that it will be possible to ascertain through circumstantial evidence in the sessions whether or not (by generally accepted cultural standards) Jane's father has/has not treated her in an aloof or neglectful way. An ancillary assessment (again, on the basis of evidence in the sessions) would also need to be made as to whether there are grounds for believing Jane is exaggerating, deluded, neurotically oversensitive etc.. The latter would be relevant factors in our assessment of her father's conduct (and also of Jane's general psychological state). It might also be possible to interview Jane's father separately to evaluate his attitude and behaviour towards his daughter. Let us assume that there were clear grounds for concluding that Jane's father either had or had not treated her with aloofness and neglect. α_1, of course, only concerns the former possibility. If he had treated her in that way then the case-study 'test' consists of evaluating the causal relevance of that behaviour for Jane's distress. We cannot manipulate the independent variable and the assessment is, of course, retrospective, preventing us from implementing experimental controls. These are, by

ideal standards, definite methodological deficits which might lead many to reject this informal method of testing outright. However, we cannot escape the practical realities of testing, and we have seen that there would be substantial liabilities of using experimental formats (based on the SMCR) too. Given the practical realities, can it justifiably be maintained that the experimental mode of testing α_1 is superior to making a case-study judgement? For those who are prepared to accept nothing less than conclusions validated to the very highest standards the case-study judgement will not be acceptable. However, for those who adopt that lofty stance the experimental tests of α_1 would almost certainly have to be rejected too (for the reasons given in 5.5 and 5.6).

In spite of the acknowledged potential liabilities, there are some redeeming features of making an informal judgmental test of α_1 in GPC sessions.

Firstly, this method of testing circumvents some of the principal problems that were identified for the testing of α_1 by the experimental applications of the SMCR:

> (i) The problem which the synchronic application encountered of not being able to operate directly with a single individual is obviated, because a single individual (Jane herself) is used when testing in a GPC session.
>
> (ii) The principal problem of the diachronic application - namely, that of finding test-subjects sufficiently similar in characteristics to Jane and her father - is obviated because in the GPC sessions we use Jane (and possibly her father too).

Secondly, assuming that Jane's father did behave in an aloof and neglectful manner towards his daughter, there might be strong evidence to support the conclusion that it was positively causally relevant for her distress, even though the 'test' for this amounted to nothing more than a set of informal judgements. Various factors would strengthen the reliability of the judgement: for example, if the intensity of the father's behaviour was considerable; if there was no indication that Jane was oversensitive, exaggerating etc.; if circumstantial evidence pointed to Jane

being distressed specifically about having been treated in that way by her father; and if alternative explanations of Jane's distress could be eliminated. Under favourable circumstances the grounds for concluding that Jane's father's behaviour was or was not causally relevant for her distress might be strong by rational, albeit informal, standards. Considerable dependence would, of course, have to be placed on background knowledge.

CHAPTER 6

THE SCIENTIFIC LEGITIMACY OF SEEKING INDIVIDUALISED KNOWLEDGE - A RESPONSE TO CARTWRIGHT

6.1 INTRODUCTION

The aim of this chapter is:

(i) To respond to a criticism from Nancy Cartwright;

(ii) To discuss some of the issues raised by Cartwright's criticism.

In a set of written comments to a draft version of chapter 5 Cartwright said:

> "There is...one major standard objection that I think would be raised; and I think you will either have to answer it or reorganise your thinking in some way to deal with it. The objection is: what you are doing here is accusing scientific method of being incapable of doing something it was never meant to do in the first place: to provide a test for the correct explanation of individual happenings. Science, after all, so the line goes, is meant to establish general claims, not claims about individuals. And the standard statistical methodology is, indeed, 'doing this'. This point of view goes back at least [to the] Methodenstreit. Could history or political economy be a science? The argument was that it could not be, because it studies individual happenings in all their individual peculiarity, whereas genuine science aims to establish universal connections between repeatable features (or perhaps it's better to think of them as recurring features).
>
> More recently, in the literature on causality, the distinction has been clearly marked by a number of authors: [for example] I talk about generic vs. singular causal claims....."
>
> (Cartwright - personal communication[39] - 21st December 1995)

I take the core objections in the above passage to be:

C1: "scientific method...was never meant to...provide a test for the correct explanation of individual happenings" (ibid.); and

C2: "science...is meant to establish general claims, not claims about individuals" (ibid.)

I agree that if propositions C1 and C2 can be upheld then the significance of my criticism of experimental ineffectiveness (in sections 5.5 and 5.6) will be weakened or undercut. α_1 is an example of the type of individual causal claim that Cartwright has in mind. Moreover, the synchronic and diachronic experimental applications of the SMCR are representatives of (applied) scientific method. In chapter 5 I operated with the assumption that it was entirely reasonable to evaluate how well those experimental applications could be applied to test α_1. But

if C1 and C2 are true then there was an error of inappropriate application on my part in even conceiving α_1 (or similar hypotheses) as suitable candidates for being tested by scientific (specifically, experimental) means.

I reject the charge of inappropriate application. In 6.2 I shall provide a counterexample to C1 and C2. In 6.3 I shall provide further reasons for rejecting them and shall try to throw light on some of the confusions which underlie their continued espousal. Finally, in 6.4 I shall argue that, beyond the mere legitimacy of seeking experimental tests of individualised hypotheses, we are entitled to view applied scientific method critically insofar as it fails to provide effective tests of such hypotheses.

6.2 A COUNTEREXAMPLE TO CARTWRIGHT'S OBJECTION

A simple way to refute C1 and C2 would be to provide an unambiguous example in which science did attempt to establish a claim about an individual happening (contrary to C2), and in which (applied) scientific method was used to test the correct explanation of that happening (contrary to C1). The example I shall use is the extinction of the dinosaurs and, specifically, an hypothesis about the possible cause of this event. I have chosen this example because it is dramatic (and entertaining), but it would have been possible to have made the same methodological point by using examples from other fields such as archaeology, pathological medicine, geology, planetary science or forensic science.

I assume that it is legitimate to regard the extinction of the dinosaurs as a singular event or "individual happening" (ibid.), albeit one of large scale. It may be objected that the event covers a biologically related group of animals and is therefore a generic and not a singular event. But this objection is unfounded. Although the animals in the group are ancestrally related their extinction, qua event, is a 'one-off' and is one which, in palaeontological terms, seems to have occurred in a very short space of time.

There is good fossil evidence that dinosaurs lived and evolved for over 150 million years but then abruptly died out approximately 65 million years ago together with a large number of other archaic groups (for simplicity we shall focus on just the dinosaurs). Fossils of all these animals are to be found in rocks of the Cretaceous period or earlier, but none have ever been found in rocks belonging to the subsequent Tertiary era (the boundary between the Cretaceous and Tertiary periods, dated to about 65 million years b.p. and in which these fossil types come to an end, is called the K-T boundary).

What could have caused the dinosaurs' extinction? Before considering some of the theories proposed we should note that in searching for 'the cause' of their extinction researchers had in mind some principal overriding novel factor or factors which brought about their apparent sudden demise. That is, 'the cause' sought has to be appreciated as something which contrasts with the presumed ongoing normal background circumstances (e.g. habitats which had sustained dinosaur survival and reproduction for millions of generations), and which interfered

with it so as to bring about the explanandum event (i.e. dinosaur extinction). The event(s) which might serve as the explanans is therefore fixed by our pragmatic interest. It would not serve the purpose of the kind of explanation we are seeking to be told that, for example, the dinosaurs became extinct because their hearts stopped beating, or because the electrical activity in their brains ceased, even though this would be true. The general point about the pragmatic or interest-relative way in which we conceive of 'the cause' of an explanandum-event has been brought out clearly by Hart and Honoré [1956] (quoted in Hewstone [1989] 4) and by Mackie ([1974] 34-35). Indeed, it is implicit in the latter's notion of a "causal field" (ibid.) - i.e. the 'field' of normal ongoing circumstances against which some intervening event strikes us as being 'the [significant] cause' because it is relevant to our interests.

Various theories for the dinosaurs' extinction were suggested including the rise of egg-eating mammals, major climate change, large-scale volcanic-like activity (Flood Basalt Eruptions), and some kind of cosmic catastrophe. The latter was proposed by Luis and Walter Alvarez of the University of California at Berkeley in 1979 when, during routine sampling of clays of late Cretaceous age in Italy, they discovered an iridium-rich layer of rock at the K-T boundary. Iridium is an element which normally occurs only in very low concentrations in the Earth's crust but is found in concentrations several thousand times that in meteorites and asteroids. The iridium concentration in the Italian clay was about 30 times background Earth-crust levels. Originally, Alvarez and Alvarez proposed that the iridium had settled as fallout from a 'nearby' supernova explosion but, when it was pointed out that this would be implausible on astronomical grounds, they changed their hypothesis to that of a massive meteorite impact with the Earth. They conjectured that if such an event had taken place the meteorite could have disintegrated or vaporized upon impact, engulfing the Earth in an iridium-rich cloud which later settled to form the layer they had discovered. Such an event could also, conceivably, explain the mass extinction at that time (65 million years b.p.) - for example, by blocking out sunlight for an extended period, or otherwise seriously disturbing the ecosystem.

So here we have a singular hypothesis - i.e. that a single giant meteorite impact was the crucial event which led to the extinction of the dinosaurs approximately 65 million years ago. One obvious way of increasing the severity of

the test of this hypothesis was to see if the iridium-rich layer which was a conjectured consequence of it was a world-wide phenomenon, and was not located only at the Italian site at which Luis Alverez had first discovered it. This is because if 'fallout' from a meteorite impact had caused the iridium-rich layer, and that impact had also been capable of wreaking catastrophic ecological damage, the geographical distribution of the fallout would have had to have been considerable. Throughout the 1980s attempts were made to answer this question and it was indeed found that there was an iridium-rich layer at the K-T boundary world-wide.

In the late 1980s attention focused on the Caribbean as a possible site for a massive, ancient meteorite impact: tektites (small blobs of initially molten rock ejected from the site of an impact) were found in Haiti, and in 1990 the glassy cores of samples of these were independently dated to 65 million years b.p. ('Astronomy' magazine July 1991, page 33). More recently a possible impact crater of the right size and age to fit the exterminating-meteorite hypothesis has been found on the Yucatán peninsula of Mexico (ibid.).

Although we can, of course, never be absolutely certain that the meteorite hypothesis is correct or that the Yucatán feature is the remnant of the very impact that caused the late Cretaceous extinction, both of these hypotheses are currently the most favoured explanations – a major turn-around since the 1970s. The idea that single catastrophic events could have had and in fact have had a major influence upon the evolution of life has, in general, been strongly resisted by modern biologists and palaeontologists who tend to favour slow, gradual changes.

The significance of the above example for our (methodological) purposes is that, I claim, it stands as a counterexample to the core claims in Cartwright's objection (i.e. C1 and C2). Patently, not only were technical methods of science used to collect data relevant to the meteorite-impact hypothesis (e.g. iridium concentrations in rock; dating techniques), but scientific methodology (in the logico-epistemological sense) was used to assess how the evidence which was available or could be obtained could best be used to test the hypothesis at issue. That hypothesis – namely, that a giant meteorite impact was 'the cause' (in the sense indicated earlier) of the dinosaurs' extinction – referred to an "individual happening" (Cartwright ibid.). The tests were directed at the crucial question of whether a singular event of the type in question had occurred in the right place at the

right time, otherwise it could not – even in principle – serve the explanatory purpose the researchers had considered for it.

It can be of no benefit to an objector to point out that the proposal of the explanatory hypothesis, and especially the carrying out of the tests, made use of a great deal of generic knowledge (for example, in the form of background knowledge) since that was not at issue and had not been denied. What was at issue is whether science and its methods ever seek to test hypotheses about individual occurrences or individual causal influences. It is felt that the example shows that at least sometimes they do and that, moreover, singular occurrences or influences can not only be interesting but theoretically important. (The above reconstruction of the story behind the search for a dinosaur-exterminating meteorite used the following references: 'Astronomy' magazine, July and September 1991 and October 1995; Norman [1985] 194-197. For a readable and up-to-date account see Alvarez [1997].)

6.3 CLARIFICATIONS RELATING TO C1 AND C2

If the counterexample of the previous section is not felt to provide sufficiently strong grounds for rejecting C1 and C2, the following consideration adds further weight.

For many years researchers in psychology and psychiatry have sought to develop a more rigorous approach to the study of the individual than is provided by the uncontrolled case-study. Hersen and Barlow ([1976] 22-28) review some of the key figures in this attempt to develop a methodology for the scientific study of the individual. These include G. W. Allport, M. B. Shapiro, D. T. Campbell, J. C. Stanley and J. B. Chassan:

> "Allport argues most eloquently that the science of psychology should attend to the uniqueness of the individual (e.g., Allport, 1961, 1962)....
>
> As early as 1951, Shapiro was advocating a scientific approach to the study of individual phenomena, an advocacy that continued through the 1960s (e.g., Shapiro, 1961, 1966)....
>
> Unlike Allport...Shapiro went beyond the point of noting the advantages of applied research with single cases and began the difficult task of constructing an adequate methodology....
>
> [Shapiro's]...important contribution...was the demonstration that independent variables in applied research could be defined and systematically manipulated within a single case, thereby fulfilling the requirements of a "true" experimental approach to the evaluation of therapeutic technique....
>
> ...Campbell and Stanley...arrived at similar conclusions [to Shapiro] on the possibility of manipulation of independent variables and establishment of cause-effect relationships in the study of a single case....
>
> [Chassan, 1967]...emphasised the various statistical procedures capable of establishing relationships between therapeutic intervention and dependent variables within the single case."
>
> (Hersen and Barlow op. cit.. N.B. Details of the references mentioned by Hersen and Barlow are provided in 'References'.)

It is therefore not true that scientific investigators are not interested in the individual, or that scientific method is inappropriate for being applied to the individual case (see also the very end of section 5.5).

Given the above, consideration needs to be given to why objections of the kind Cartwright raises[40] persist (she describes these objections as "standard" – ibid.). At the heart of the objection seems to be the belief that science has a legitimate interest only in generalisations (or generic knowledge). Mirroring this is the idea that scientific method is suitable only for acquiring such generic knowledge

or testing generalisations. Whereas no sensible person would deny that generalisations are of fundamental importance to science and its methods (see e.g. Reichenbach [1951] 5), it is the converse claim of the illicitness of individualised knowledge for science that is so puzzling. Why would anyone seriously maintain this? Even though I do not pretend to know the full answer, I do have a suggestion. Namely, that this (regrettably) pervasive view stems from certain confusions. I shall now discuss various issues connected with these confusions as a series of points:

(1) Firstly, I think there is confusion in the use of words and phrases such as 'singular', 'singular hypothesis', 'individual', 'individual happening' etc.. What the proponents of C1 and C2 might have in mind by such terms are strictly unique occurrences (or hypotheses about these). This, with an important qualification that I shall make in point (2), below, is *not* what I have in mind when I have defended the position that science and its methods are (or should be) applicable to singular or individual cases. Sometimes I have meant by the term 'individual' a single human person; and I have generally meant by the term 'individualised (or singular) hypothesis' an hypothesis about a single person. It is obvious that a single person can be the site of multiple or repeated effects (e.g. it is possible that Jane was the recipient of aloof and neglectful behaviour by her father on many occasions). Similarly, if we conceive of a single human individual as a physical system (be that stochastic or deterministic), that system can exhibit recurrent properties of a certain type (e.g. heartbeats, electrical impulses, states of mind of being distressed at p, pleased that q etc.). Even when the term 'individual' (or 'singular') is used to refer to some specific event (e.g. Jane's mental state of distress on a particular occasion; or Tom's being subjected to an aggressive interrogation on a particular occasion – see section 3.32) we do not typically intend a strictly unique occurrence (with the qualification to be made in point (2), below), but one which has significant similarity to others with which we are already familiar (e.g. other cases of distress; other cases of children intimidating children).

Overall, we need to distinguish carefully between: (i) a sense in which events in personal, social or cosmic history are unique (see point (2)), and which is *not* the focus of scientific interest; and (ii) a legitimate concern which science and its methods may have in certain events which occur only once.

In addition, as noted, we should not become confused by the language of speaking about individual persons or happenings, or hypotheses about these. Often, the intended meaning is not some strictly unique attribute but, rather, attributes of individual persons or single events which are shared by other persons or events (so that the referent is, in fact, a member of a class).

(2) We may now, as the second point, turn to the qualification mentioned above. When speaking about phenomenally knowable events in the world there is a sense in which *every* event is unique or 'singular' (i.e. has never occurred before with the same space-time identity, or is of a novel type). Also, if we are studying a physical system S then, almost inevitably, there will be changes to this system with time, however slight the difference may be. S at time t will not be identical to S at time $t + \Delta t$ (and the *exact* set of properties at any one given time may never recur). It is true, for example, that the event referred to in hypothesis T1 (see page 87) is unique in the history of the world (or, at least, we have very strong reasons for believing this to be the case). Tom is a unique human being and the exact circumstances of the event in the playground when he was intimidated by Dan and his accomplices never occurred at an earlier past time, has never recurred since then, and will never recur in the future. Uniqueness of this kind is of the same type that we find in history (e.g. the moment that the young Beethoven first met Mozart) or in social and economic events (e.g. the series of strikes in Britain in 1978 known as 'the Winter of Discontent'). The events of interest to historians, sociologists and economists are often large-scale, because they typically deal with collective or mass phenomena (but note that historians also frequently focus on individual events in a person's life). In contrast, psychotherapists are interested primarily in the individual and in significant occurrences in the life of that individual. However, both the large-scale events of history and society, and the small-scale ones in the life of an individual, are unique in the sense outlined. The proponents of C1 and C2 are correct in their implication that science is standardly not interested in unique occurrences, *qua unique*. However, occurrences that are unique in the sense characterised may have attributes in common with other occurrences. Similarly, an individual or system that is changing from one moment to another may share attributes with its former states, or may exhibit recurrent features. This allows, at

least *prima facie*, for the uniqueness objection to be obviated, since many occurrences, individuals or systems in the human and historical domains will exhibit non-unique or recurrent properties permitting, in principle, their study by methods which require such properties to obtain. If they cannot be studied in an exacting scientific way then we should attempt to understand the reason(s) for this. It could be that the uniqueness or individuality of the events/systems involved has little or nothing to do with the epistemic failure, and that Cartwright's objector's have provided us with a red-herring. The reason(s) for the comparative failure of being able to develop an exacting science of the human and social domains may reside elsewhere – e.g. with the *complexity* of the systems involved (Rosenberg [1988] 10), or with the inherent inductive limitations of current scientific techniques.

A further important point, which will be elaborated as item (4), below, adds further weight to the irrelevance of the 'uniqueness' objection. This is (I maintain) the fact that the uniqueness which Cartwright's objectors impute to history, society and individual human psychology applies equally to the systems studied in the physical sciences. If this is correct, then the objection that it is the uniqueness of events or systems that is prohibitive for those events/systems being scientifically studied could be refuted by the fact that events and systems studied in physics and chemistry are also unique, but that this has not prevented these fields from being developed as *bona fide* sciences. Before discussing this, I shall articulate as item (3) a legitimate critical point about testing causal claims which is implicit in Cartwright's objection.

(3) Earlier, it was acknowledged that there was a sense in which if events were strictly unique the proponents of C1 and C2 would be justified in saying that science would not be interested in them or that its methods would not be competent to deal with them. This point (my third main one) needs to be elaborated. I shall focus on causal knowledge, though I think the general point would also apply to descriptive knowledge[41]. Our rational understanding of causation tells us that causes cannot be known if either of the events which are potentially related as cause and effect are strictly unique occurrences. If two events X, Y were absolutely unique – i.e. there was no known X' to which X bore a phenomenal resemblance, and no known Y' to which Y bore a phenomenal resemblance - and if either X or Y, or both X and Y, each occurred only once, then we would not be able to know whether or

not they were causally related[42]. Under these conditions it would not be possible evidentially to discriminate the following propositions:

(i) X and Y are causally related

(ii) X and Y are causally unrelated (and just happen to occur one after the other)

This would also apply if X, Y resembled (respectively) X', Y' but X', Y' did not exhibit any pattern of joint recurrence.

Consequently, if we had strictly unique (i.e. truly singular) events there would be no point in attempting to apply experimental methods to them to evaluate if they were causally related. If this is the point that proponents of C1 and C2 were making, then it is correct but, as indicated earlier, it is one that no reasonable person would deny. It is, moreover, irrelevant as a critique of those such as Shapiro who aimed at developing a methodology for testing hypotheses (including causal hypotheses) about individuals. Shapiro and the other researchers mentioned (see page 182) never, in my view, believed that they were dealing with strictly unique occurrences in any sense that would not apply equally to systems in the physical sciences.

(4) Suppose I am a chemist who is studying the (causal) influence that the addition of an alcohol (-OH) group to chemical R has on the rate of reaction when R combines with S to produce, say, T plus U. That is, I wish to compare the rate of reaction of:

(A) $R + S \rightarrow T + U$

With

(B) $R\text{-}OH + S \rightarrow T\text{-}OH + U$

Assume I already have a value for the rate of reaction (A) and that I set up an experiment to measure this value for (B). I will need to measure out quantities of R-OH and of S, mix them, and then measure the rate at which these

substances convert to T-OH and U. Suppose I do this. Would Cartwright's objectors complain that this is not a suitable subject for scientific investigation or that scientific method is not appropriately applied in this case? Why not? I would defy anyone to repeat exactly the experiment for (A), but in which R-OH replaces R. (This would also apply if one just wished to repeat (A), or to repeat (B).) Given our best theories of science, general practical background knowledge, and actual skills of manipulation, it would in practice be impossible to recreate *exactly* the circumstances in which I conducted the experiment for (A), only now using R-OH instead of R. However carefully one attempted to do so, there would be some variation (apart from the intended addition of the –OH group). Perhaps a fraction of an extra drop of R-OH was added than was R in the original experiment. This is a huge difference – billions of extra molecules. Or perhaps there was a tiny difference in the temperature (i.e. the mean kinetic energy of the molecules) of the combined solutions as compared to the first experiment. Perhaps a cloud of a different density passed in front of the Sun as it shone through the laboratory window during the first experiment as compared with the second. This would change the thermal radiation received by the reaction vessel (and note that temperature is causally relevant for reaction rate).

It would be possible to multiply examples *ad nauseum* (in physics, for example, consider the practical impossibility of repeating *exactly* Eddington's famous solar eclipse experiment of 1919 to measure the bending of light rays by a gravitational field).

My point is that we have no less reason to suppose that the natural systems studied by physics and chemistry are unique (in the sense outlined in (2), above) than we do of those studied by history, sociology, economics or psychology.

(5) The view has been taken that the world, at least in its phenomenal properties, is in a continuous state of flux. It exhibits novel properties and relations from moment to moment and this, so far as we know, is most extreme or noticeable in domains such as the biosphere, evolution, human history and society, and human psychology. This claim is intellectually well-grounded through countless examples. On the other hand, it is not being claimed that everything is undergoing change. There is also a basic intellectual intuition which needs to be satisfied to the effect that there are certain underlying invariants. These invariants – which are usually

articulated in terms of laws of nature or capacities – are postulated as the means of being able to explain why things happen as they do and why there are certain regularities or patterns at the phenomenal level. They similarly counter the expectation of chaos that would follow if there was *only* continuous change and novelty. There are, therefore, two apparently opposing features of nature which need to be recognised (N.B. I take nature to include the human social domain). These are: (i) (as far as we know) the endless appearance of novelty, uniqueness, singularity; (ii) the existence of underlying invariants (or, at least, very stable constancies).

The second of these is, on a Realist construal, if anything, more problematic than the first because, whereas there is abundant and comparatively direct observational evidence to support the former, the latter involves the postulation of unobservables (i.e. laws of nature, or capacities). The rationale for the postulation of these unobservables involves, moreover, a certain kind of metaphysical thinking – i.e. the need to account for appearance and change in terms of deep, hidden structure which lies 'behind' it[43].

My fifth point is that the existence of uniqueness and change does not imply that there are no laws or regularities, or that causes cannot be discovered. So, one can both recognise the abundance of singularity in nature, and yet be intellectually justified in searching for, and claiming (with appropriate evidence) to have found, scientific laws. However, one must acknowledge that scientific laws are to some extent idealisations, abstracted from the precise phenomenal features of the circumstances in which they are supposed to hold (see e.g. Cartwright [1983] chapters 3, 4 and 6). I am not, of course, denying that there are still considerable difficulties in being able intellectually to 'reconcile' fully these two perspectives on nature. However, I do not think that this desideratum is a good enough reason for rejecting either perspective (or their joint acceptance) as legitimate.

6.4 THE NEED TO BE OPEN ABOUT THE FAILURE OF SCIENTIFIC TESTING

An important feature of Cartwright's criticism was its claim that it was mistaken of me even to expect scientific method to be applicable to the individual case:

> "[Y]ou are...accusing scientific method of being incapable of doing something it was never meant to do in the first place...."
> (Cartwright - personal communication - see Note 39)

This can be viewed as: (1) a charge of misapplication. That is, that to apply scientific method (e.g. as an experimental test) to the individual case is to misapply it. But coupled with this there seems to be a second, tacit, view. Namely, that: (2) if scientific method fails effectively to test hypotheses about individual occurrences that should not be taken to reflect badly on scientific method (since, according to '(1)', scientific method was never intended for such an application). The emphasis in this section is intended to be on '(2)'. However, before turning to it, a little more needs to be said about '(1)'.

There clearly is a sense in which it would be possible to misapply scientific method or, more specifically, a canonical scientific testing format. This point can be illustrated by reference to statistical tests. In statistical testing there is a clear acknowledgement that different tests (or test statistics) have different applications, and that each is constrained by factors which limit that applicability (Miller [1984] 56-66). For example, parametric tests (such as the Z- and t-tests for independent samples) operate with samples which can be assumed to have been drawn from a normally distributed population (op. cit. 61). If, in designing an experiment or in carrying out a statistical analysis of the data we used a parametric test when in fact the parent population was not normally distributed that would constitute a trivial misapplication of the test in question, because we would have broken the ground rules for its application. In such a situation we would not expect to draw any deep methodological or epistemological lessons from the failure of the test. However, I have maintained (at the end of section 5.5) that when I analysed whether the synchronic and diachronic experimental applications of the SMCR

could test α_1, this did not constitute a misapplication, trivial or otherwise. My reasons were as follows.

In the case of the synchronic format, even though that format cannot work directly with a single individual, conclusions about individuals are regularly reached on the basis of the results of group-comparison (i.e. synchronic) studies. This typically involves making a judgement about the individual case given the generic conclusion. Such judgements are commonplace: e.g. assessing the likely effect that administering a drug D will have on an individual M, when the effect of D has only been tested in group-comparison clinical trials. In such a situation the 'test' of the individual hypothesis consists in forming a reasoned opinion about what would happen if D were administered to M, in which the generic result is used as a critical datum in forming the opinion. Given that a group-comparison conclusion can, therefore, be used in inferring what is believed to be the likely outcome for an individual, it cannot be wholly inappropriate - or deemed a misapplication - to enquire into whether the group-comparison format could have been applied directly to test the individualised hypothesis and, if not, why not.

In the case of the diachronic format, ample evidence has already been supplied to show not only that this application is not a misapplication, but that being able to test certain individualised hypotheses is one of the format's specific and intended design features.

Let us now turn to a discussion of point '(2)'.

The danger inherent in point (2) is an unwarranted defensiveness about science and its methods. The objectors Cartwright refers to maintain both that scientific method was never intended for testing singular hypotheses, and that the socio-historical disciplines are to be distinguished from the natural sciences by virtue of being replete with claims about non-recurrent events. Confronted, then, with the apparent fact that the socio-historical disciplines, unlike the natural sciences, have failed to develop into *bona fide* sciences (Rosenberg [1988] 6-13), a ready-made explanation is at hand for this difference. Namely, that it is to be expected, given that applied scientific methods are not meant for testing singular hypotheses. This perspective has the capacity to exonerate scientific method from responsibility for the comparative scientific underdevelopment of the social and

historical disciplines. On this view, the 'fault' lies not with science or its methods, but with the intrinsic interests of those working in the human, social and historical fields. These interests (e.g. singular historical events; the mental states and motives of individuals) are regarded as wayward from a scientific point of view.

An alternative to the above can, however, be provided. On this view the above perspective reflects, at most, only a very limited portion of the truth. According to the alternative, emphasis should be placed on the limitations or failure of science and its methods. When considering the relative scientific underdevelopment of the social and historical disciplines what needs to be highlighted is not the intrinsic interests of those working in these fields, but the inductive incompetence of scientific method for being able to turn them into *bona fide* sciences. On this account it is applied scientific method that is wanting.

There will be cases in which scientific method should, rightly, *not* shoulder the blame if effective testing fails. As indicated earlier, a method should not be blamed if it is trivially misapplied (see page 189), or if the hypothesis to be tested is not formulated with sufficient clarity, or if the constructs serving as independent or dependent variables are empirically meaningless etc.. However, given that basic adequacy conditions are met, if effective testing is still not achieved (or achievable) it is appropriate to ask if there is something more fundamental awry. I have conceived of scientific method in the present context primarily as a set of applied methods – i.e. as a set of inductive instruments. It is certainly legitimate to enquire into why these might be limited in their effectiveness in certain ways, or even fail completely. Of particular interest would be *patterns* of inadequacy or failure. For example, it might be that applied scientific methods show especial weakness (or even complete failure) for testing hypotheses about mental states individuated by (intentional) content. If so, *why* should this be, given that most psychologists and philosophers of mind are nowadays physicalists, and so have no objection to those states being conceived of as physical (see e.g. Greenwood [1991] 6)? Is it because of pragmatic restrictions on being able to carry out the test; is it because of some limitation of the logic (i.e. inferential capability) of the test; is it because of the empirical inaccessibility of the variables; or what? The point is that we should be prepared to look upon applied scientific method as potentially capable of failing, even in a systematic way. We have focused so far on individualised hypotheses in the social and historical domains, but the moral is one that is relevant

for applied scientific method in general, regardless of whether the hypotheses are singular or generic, or the domain social or natural.

Nothing anti-scientific should be read into the above attitude. It is in keeping with a critical epistemological outlook which is prepared to review the methods of scientific enquiry (as well as the natural inferential and inductive capacities of human cognition) from a rationalistic standpoint. A basic lesson is that we should be prepared to accept that there might be systematic or localised failure of a deep-rooted kind with regard to the ability of experimental methodology to test certain categories of hypotheses effectively. I am not claiming that this will turn out to be the case, but I am opposed to a scientistic dogma which refuses to countenance it because of a misplaced allegiance to science and its methods. Equally, I am not advocating that we should not try to adapt and apply methods of scientific testing to the social domain (on the contrary, I think this needs to be undertaken vigorously). Rather, what I have in mind is a preparation for the possibility that, at least in the social sphere, a highly optimistic view of the capabilities of rigorous scientific testing may have to be abandoned. But, more than this, the critical attitude I am advocating endorses:

> (i) Taking the first steps towards actually searching for possible domains in which effective scientific testing (especially by experimental formats) breaks down;
>
> (ii) Documenting the nature and extent of the failure;
>
> (iii) Attempting to elaborate an epistemic theory for why failure occurs (especially if it seems that the reasons are so fundamental that future attempts to resolve the hiatus may be in vain).

Of course, it would be unwise to stipulate dogmatically that there could never be successful experimental testing of an hypothesis which extant experimental methods appear incapable of testing effectively. It is impossible to predict how testing methods will be developed in the future. On the other hand, it may be possible to provide good reasons (theoretical or practical) for why we would not expect certain hypotheses to be effectively tested experimentally, even in the

very long term. Exactly how applied scientific methodology does pan out remains to be seen, but we should no more expect success as a foregone conclusion than rescind on using resourcefulness in an attempt to make it work.

CHAPTER 7

PROBLEMS OF JUSTIFYING FP KNOWLEDGE-CLAIMS

7.1 INTRODUCTION

Given that such great reliance needs to be placed on FP conclusions both in everyday life and in psychotherapeutic counselling (GPC), some discussion of how FP hypotheses are tested and justified is in order.

A distinction has already been drawn (in 2.3 and 3.2) between: (i) being able to provide a global justification for FP claims; and (ii) being able to validate individual FP claims (be they singular or generic) to a very high standard (i.e. akin to that required for scientific demonstration). I think that whereas '(i)' can be achieved (see 7.22), the prospects for '(ii)' are generally poor.

Principal, though not exclusive, attention will be given to examples of *causal* FP hypotheses, and the problems of validating these. This is because of the focus by Grünbaum and his supporters (e.g. Erwin) on causal hypotheses. It should, however, be borne in mind that the problem of justification for FP knowledge includes hypotheses that are not overtly causal but descriptive, and that in some respects these are primary. For example, prior to considering any possible causal influence that Mary's mental states (e.g. belief B, desire D or fear F) might have on her action A, we need to know whether Mary is in these states and, if so, what is their intentional content(s). There is an obvious epistemological problem in being able to acquire reliable descriptive knowledge of the mental states. In this thesis I have nothing specific to say about the descriptive problem and, because of this, the analysis in this chapter is not intended to be systematic. Furthermore, even when causal hypotheses are discussed the examples are limited in scope.

A further preliminary issue concerns the nature of FP-reasoning as an inductive capacity. This involves such questions as: how, as a matter of fact, does FP-reasoning operate (i.e. how does it 'work'); what inferential strategies does it utilise; how are the inferential strategies either innately built-in to human cognitive functioning or learned; how are causal relevance discriminations made by means of it; etc.? The view has already been taken (see pages 38, 153, 171, 228-229) that we currently have a very imperfect understanding of these matters. The inductive strategies by means of which we are capable of understanding ourselves and others are themselves not well

understood. Significant attempts have been made by philosophers and psychologists in recent years to shed light on this interesting topic. For example, there are currently two main (and, to a large extent, competing) views within the philosophy of psychology regarding the method by which interpersonal understanding is achieved: namely, the 'Theory Theory' and the Simulation Theory (Davies and Stone [1995] 1-44). In addition, social psychologists have taken a more empirical approach to the kinds of attributions that we make of each other and have attempted to formulate principles by which these attributions are made (this is usually called Attribution Theory – see e.g. Hewstone [1989]).

As philosophers, our primary concern in this field must be with the methodological and epistemological issues. For example, we need to suggest plausible accounts of the kinds of inferential strategies that would have to be used as part of FP-reasoning so that the kinds of conclusions that we have good reasons for believing are reliably reached by its means can in fact be reached by it.

However, I believe that a naturalistic (or empirically-guided) approach is essential in any attempted reconstruction of FP-reasoning capabilities. We *do* need to analyse such capabilities in logico-methodological terms (and this implies conceiving of them essentially as the implementation of *a priori* inferential strategies). On the other hand, if a reconstruction is to be even approximately accurate it must take into account empirical facts about, for example, how human cognition works, the kinds of real-life situations in which FP inferences are made, how the inferences are conducted in those real-life situations etc.. 'Ecological validity' is, therefore, crucial in attempts to understand and reconstruct FP-reasoning capability. This is something I have emphasised in my criticism of Grünbaum's sketch of FP causal inference for the case of insults (see 7.3).

Overall, since we cannot possess purely *a priori* knowledge of how FP-reasoning works, our attempts at reconstructing FP-reasoning capability must consist of theories or models. These should include elements that allow us to explain, logically, how the requisite inferences could be made, but in combination with any relevant empirical facts which constrain how they are in fact made. Various authors have already provided rudimentary models (or sketches for models) of this kind (e.g. Hopkins [1991]

88-96; Grünbaum [1993] 163-166; Wellman [1990]). In this chapter I shall confine detailed discussion to aspects of Grünbaum's account for the specific case of insults.

It is important to appreciate that these various models are all predicated on the *assumption* that FP-reasoning as employed in everyday life does to a significant degree issue in epistemically justified - or true - FP conclusions. The primary function of the models is to show *how* the FP knowledge that we believe (or assume) we possess could be arrived at, and how it could be justified. This function of the models does not satisfy the task of providing basic justification for any given FP hypothesis. However, it is the latter that the sceptic of FP knowledge challenges us to provide. It is important to keep this in mind, because whereas the various models do possibly lend some background plausibility to the antisceptical case (see 7.22), they may not answer in a completely satisfactory way the core problem raised by the sceptic, which is: "why (i.e. on the basis of what rational utilisation of evidence) should we accept the specific FP claim, P, that is being asserted?".

Turning more squarely to the problem of validation: as already indicated, the prospects for validating FP hypotheses to high standards is currently poor and likely to remain like that for the foreseeable future. It may even be a permanent feature of limitations on scientific validation, though it would be unwise to make predictions about the distant future of scientific testing. Assuming FP ontology to be genuine (which is a controversial point – see pages 73 and 199 of this thesis), it seems that the problems of rigorously testing individual FP claims are largely pragmatic. That is, the apparatus (physical or logical) of experimental tests cannot be brought to bear effectively upon the relevant phenomena. An important reason for this is their often transitory nature. Hopkins, for example, remarks that:

> "...motives constantly vary, in response to need, experience, and thought, and so rarely satisfy the same description from instance to instance."
> (Hopkins [1991] 129; footnote 21)

Even if, hypothetically, the relevant phenomena could be 'held still' for long enough it might still not be possible to apply an experimental test effectively (see chapter 5), or to manipulate the independent variable. Even the physical identification

of an FP mental state with a brain state – something that has been assumed for present purposes as a metaphysical premise – has no significant prospect of being empirically characterised in the foreseeable future (see e.g. Baron-Cohen [1995] 22-23; Wilkes [1991] 26-27). Yet any fully independent experimental test of, for example, the causal relevance of a motive for an action would seem to require the empirical characterisation of that motive as a brain state, its empirical detection, and possible manipulation.

In the remainder of this chapter I shall proceed as follows:

In 7.2 I shall present the problem of justifying FP knowledge in terms of two contrasting perspectives on the matter – scepticism (in 7.21) and anti-scepticism (in 7.22). Whereas there are difficulties for both perspectives antiscepticism, in my view, deserves to be favoured.

In 7.3 I analyse a sketch made by Grünbaum for how the causal relevance of insults for introspectively experienced emotions can be inferred through FP-reasoning. I criticise Grünbaum's account because it is unlikely to be empirically accurate. My general point is that attempts to model FP-reasoning (which is, in effect, what Grünbaum is doing) must be adequate *both* logically *and* empirically.

In 7.4 I present an alternative model of my own for inferring causal relevance in the case of insults. I maintain that this model is in some respects superior to Grünbaum's.

In 7.5 I discuss some criticisms of my model made by Nancy Cartwright and Gabriel Segal. I (currently) have no satisfactory answer to these criticisms, and the issues raised by them go beyond the scope of what can be adequately dealt with in the remainder of the thesis. The chapter therefore closes with an admission of the need for further research on this topic.

7.2 SCEPTICISM AND ANTISCEPTICISM WITH REGARD TO FP KNOWLEDGE

7.21 SCEPTICISM

Amongst those who are sceptical about FP knowledge we can detect differences in the nature as well as in the degree of their scepticism. Regarding nature we can, for example, distinguish between those who believe that the ontology of FP is theoretically unsatisfactory or even fictive and those who, while accepting the ontology, present serious doubts about our ability to make reliable epistemic discriminations about the kinds of states or causes FP posits in real-life (and experimentally uncontrolled) settings. Eliminativists such as Paul Churchland reject the ontology of FP as bogus and hence fall into the category of what we might call the 'ontological sceptics' of FP. For example, Churchland says:

> "Eliminative materialism is the thesis that our commonsense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience.
> (Churchland [1989] 1)

And;

> "...the concept of a witch is an element in a conceptual framework that misrepresents so badly the phenomena to which it was standardly applied that literal application of the notion should be permanently withdrawn. Modern theories of mental dysfunction led to the elimination of witches from our serious ontology.
> The concepts of folk psychology - belief, desire, fear, sensation, pain, joy, and so on - await a similar fate, according to the view at issue [i.e. Eliminative Materialism]."
> (Churchland [1988] 44)

On the other hand, in personal discussions with me John Worrall has repeatedly argued for our inability to know with reliability even simple FP attributions or putative causes even though he does not appear to reject FP ontology. Also, Edward Erwin in at least one published passage (Erwin [1993] 446) does the same. This makes Erwin and Worrall what we might call 'epistemological sceptics' of FP.

The depth of scepticism with regard to FP also varies. All of the aforementioned sceptics are very strong. In this thesis I shall not be attempting to provide a response to Eliminativism and so shall not attempt to provide a response to ontological sceptics like Churchland. But even amongst the epistemological sceptics mentioned the scale and depth of the doubts they raise are considerable.

For example, Erwin does not allow it to be taken for granted that a person's desire to (say) drink a glass of water is causally relevant for his subsequent act of going to a tap, filling a glass with water and drinking without the putative causal connection being established (Erwin op. cit.). Also, in personal discussions as well as in written comments to draft versions of this thesis John Worrall has raised sceptical objections of a comparable kind. For example, in response to an example which I had planned on using ('Sue is angry at the 9 a.m. train [which she was expecting to catch] being late') Worrall pointed out that, without adequate validation, I cannot be allowed to maintain that what Sue is angry about is the lateness of the 9 a.m. train (this is even if she claims that that is what she is angry at). Worrall argues that Sue could be mistaken about the object of her anger. Perhaps she believes that she is angry at her train being late but is actually angry at something else - for example, at her father with whom (say) she had an argument that morning. In a written comment to an analogous example (in this case anger at having been insulted) Worrall says:

> "You *conjecture* that your anger is anger-at-being-insulted. The question then is whether the conjecture is *correct*/rationally justified. ...there is a real issue as to whether that's what your [anger] is really about."
>
> (Worrall – personal communication, January 1997; emphases in original)

I *do* accept that Worrall's objection is intellectually legitimate (similarly for Erwin's). Indeed, it is because of objections such as these that the problem of providing a fully adequate justification of FP knowledge is so great. On the other hand I do not think that strong (epistemological) scepticism with regard to FP knowledge (whether of oneself or others) is viable as a general position. (I am making this claim relative to the assumption that the ontology of FP is acceptable.) I shall present some reasons for this in 7.22. First, let us characterise some of the forms that

(epistemological) scepticism with regard to FP knowledge could conceivably take.

(a) Consistent and complete scepticism

According to this brand of scepticism one does not place any credence in one's ability to know in commonsense terms one's own mental states or those of others, nor their causes, nor their effects. If one were consistent in embracing this form of scepticism it would seem that one would be compelled to abandon thinking, as commonly conceived, since one would be disallowed from relying on what *appeared* to be the content of one's own thoughts or beliefs. After all, one could be deceived about one's own beliefs. Similarly, one could not rely on what, by commonsensical accounts, one took to be the objects of one's emotions (e.g. that one loved one's children or hated acts of cruelty). Also, since these doubts would extend to the knowledge of the mental states of others and to the potential causal role of those states in their actions, one would be compelled systematically to resist attempts to understand other people in FP terms or to communicate with them (e.g. regarding what they thought, believed or felt). The appropriate lifestyle for a sceptic of this kind might be to remain frozen in a catatonic-like state of thoughtlessness, and to be entirely uninvolved in any form of social communication.

(b) Inconsistent scepticism

A sceptic of FP knowledge of this type has double intellectual standards. Typically, he will point out with considerable care and acumen why some FP knowledge-claims which most people would take for granted (such as that Sue *really is* angry at the 9 a.m. train being late, when she says so) cannot be taken at face value for reasons which may be quite compelling. Having done this (and possibly leaving us in a situation in which we are unable to demonstrate our initial belief), the inconsistent sceptic then fails to apply the same lesson to himself and his situation. Thus, for example, whereas he may fervently argue that there is a serious problem over whether we can reliably know whether Sue is angry at her train being late, in his daily life he

happily reaches and relies upon numerous conclusions of a comparable kind. For example, he may not hesitate to judge it as a reliable inference that (let us suppose) the woman in the bus queue on whose foot he accidentally stepped was angry at him for the act; or that the pleasure expressed by his young son at being given a new bicycle was at receiving that gift. It is to be emphasised that our objection against the inconsistent sceptic does not consist of a complaint about his having shown there to be intellectually well-grounded doubts and uncertainties in areas previously unexamined or held to be secure. Insofar as he has done the latter he has provided a valuable service and a stimulus for further research. Our objection consists, instead, of the following:

> (i) The inconsistent sceptic does not apply systematically the uncertainties he identifies, and therefore gives the appearance of being selective in an *ad hoc* way with regard to those FP conclusions which he accepts and those which he rejects.
>
> (ii) Since he appears routinely to accept, rely upon and even make predictions from many particular FP statements of the same generic kinds as those which he argues cannot be accepted at face value, he betrays an actual practical intellectual confidence in the substantial reliability of such statements, on the whole. This seems to be a form of inconsistency (or a partial refutation of his own thesis) since, if the unreliability were as severe as he claims, he ought not to be able to possess the degree of knowledge which, when not playing the sceptic, he takes himself to have.

(c) Scientific scepticism

The scientific sceptic identifies genuine reasons for doubt of the kinds we have encountered, but then adds (by implication, even if not explicitly) that the relations in question need to be justified scientifically (or to the standards of scientific epistemology). An example of an attitude of scientific scepticism can be seen, in my view, in a critical discussion by Erwin of Hopkins's account of FP-reasoning (to be

found in Hopkins [1991] 88-96). According to Hopkins, FP-reasoning involves interpretation and the attribution of motive in which the meaning (i.e. the linguistically articulable intentional content of mental states such as desires, beliefs, hopes etc.) not only gives sense to actions, but also plays an important role in our ability to infer causal relations involving those mental states. I should emphasise that I am not endorsing Hopkins's account, which is beyond the scope of this thesis to evaluate in its own right. My aim is limited to looking at some of Erwin's comments on Hopkins's paper in order to expose what I consider to be an excessive scepticism on Erwin's part. An outline of Hopkins's position is provided by the following:

> "Our most basic and familiar way of understanding the activities of persons - either our own, or those of others - is by interpreting them as actions resulting from motives, including beliefs and desires. In everyday life we do this naturally and continuously. Thus we see someone moving toward a tap, grasping a glass, and so on, and interpret this in terms of his wanting a drink, and so moving because he takes this to be the way to get one....
>
> This is a fundamental kind of psychological thinking, and one that partly defines our conceptions of mind and action. It is at once interpretive and explanatory. It is interpretive because, as such examples illustrate, assigning motives enables us to make sense of what people say and do. It is explanatory because we take the motives we thus assign to be causes within persons which prompt their actions, and which, therefore, serve to explain them."
>
> (Hopkins op. cit. 88; footnotes omitted)

Erwin presents a series of objections which have the common aim of showing that various causal FP conclusions which Hopkins believes his model of FP-reasoning is capable of providing adequate justification for are, in fact, not adequately supported by that model. The upshot is that in each case the question of whether or not there is a genuine causal connection is left begging. Erwin says:

> "Hopkins speaks of knowing what a desire is "supposed to do" (ibid.). This means what a desire *should* do if acted on intentionally. For example, he claims that a desire to get a drink, if someone acts on it intentionally, *should* produce an action of getting a drink. The basis for this claim, or even what it means is unclear. However, even if Hopkins is right about what desires should do, knowing what a desire is "supposed to do" is not sufficient for knowing its actual causal role. Do people ever act on the desire to get a drink? Does such a desire ever make any difference at all to the way people behave? These are empirical questions, even if they have obvious answers; they cannot be answered merely by articulating the content of the desire. The overlap in content between the desire to get a drink and the action of getting a drink may be grounds, then, for believing that the desire is, in Hopkins's words "supposed

> to" bring about that action, but it is not grounds for believing that it actually does so....
>
> ...It also does not follow from Hopkins's argument that the finding of motivational "sense" or "meaning" and the establishment of commonsense causal order are one and the same (ibid., 95). It is one thing to make sense of what people do by postulating motives; it is another thing *to establish* that a motive featuring content overlap with an action really did cause it."
>
> (Erwin [1993] 446; emphases in original. The references given by Erwin are to Hopkins [1991].)

Whereas I accept that it is intellectually legitimate to seek to demonstrate causal connections to adequate standards, I am also critical of: (i) the nature of Erwin's scepticism; and (ii) the implication that the appropriate standard of testing is that set by developed science. Whereas the latter standard might be desirable as an ideal, if one is making a practical prescription it is reasonable to ask only for the best that can be achieved in practice – and for the foreseeable future what can be achieved is well below the scientific ideal. I shall now discuss these issues:

(i) I am critical of the nature and degree of Erwin's scepticism about FP knowledge. Erwin does not declare himself to be an Eliminativist, and so I am assuming that his objections are 'epistemological' rather than 'ontological' (for this distinction see pages 199-200 of this thesis). Erwin says: "Hopkins speaks of knowing what a desire is "supposed to do" (ibid.). This means what a desire *should* do if acted on intentionally. For example, he claims that a desire to get a drink, if someone acts on it intentionally, *should* produce an action of getting a drink. The basis for this claim, or even what it means is unclear." (ibid.). Erwin would appear to be justified in saying that "the basis for [Hopkins's] claim...is unclear" (ibid.), if what he means by "basis" is (for example), the underlying mechanism linking the content of a particular desire to the performance of a particular action, or that linking the 'willed intention' to act upon a given desire (whatever that involves at a psychophysical level) and the action's performance. These are problems in the philosophy of mind/psychology which raise considerable difficulties at a purely theoretical level, and we certainly do not possess adequate neurophysiological explanations of them at present (see e.g. Baron-Cohen [1995] 22-23; Wilkes [1991] 26-27).

On the other hand I think it is also true that, in daily life, we all routinely

rely on the assumption that, typically, there will be a matching in content between a given desire and the kind of action that will ensue if the desire is acted on, and that we understand at a commonsense level what it means for a desire to be acted on. Consequently, I cannot agree with Erwin's remark that "what [Hopkins's claim] means is unclear" (ibid.). What further degree of clarity is Erwin seeking? It should be obvious that the frame of reference of Hopkins's analysis and discussion is that of FP (i.e. commonsense psychology), since this was stated by him repeatedly. Also, I think it is beyond reasonable doubt that even young children are competent at understanding the kinds of relations between the content of desires (and beliefs), actions, and causes and effects that Hopkins alludes to; and that they are adept at employing such relations in their enquiries and explanations (see Hopkins [1988] 38). Consider, for example, asking an infant of five or six years of age what she would like to eat - adding that on this occasion she can have whatever she pleases. Suppose she replies: "ice-cream". Suppose also that one tells her that she can have her wish, but that one then presents her with a bowl of spinach, instead. One can be fairly confident that one would receive a reply to the effect that "that's not ice-cream" or "I asked for ice-cream". Such a response is, I believe, strong evidence that the child not only understands what her desire is (i.e. to eat ice-cream), but that she understands what the correspondence of content is between the kind of desire she has and the kind of circumstance that would satisfy it (i.e. her eating ice-cream). It is also strong evidence that the infant understands at a *causal* level what sort of change her desire is "supposed" (ibid.) to bring about if it were to be realised (or acted upon), and what sorts of circumstances do not constitute such a fulfilment.

Given this substantial level of understanding and competence in FP causal locutions and inferences by even young children, why is it that Erwin seriously purports not to understand a typical, simple example ("what it means is unclear" - Erwin ibid.). Could it be that Erwin lacks a capability possessed by infants; or is it, rather, that a highly intelligent and able philosopher is to some extent feigning lack of understanding (or is there some other explanation for Erwin's alleged incomprehension)?

Erwin then says: "...even if Hopkins is right about what desires should do, knowing what a desire is "supposed to do" is not sufficient for knowing its actual

causal role. Do people ever act on the desire to get a drink? Does such a desire ever make any difference at all to the way people behave? These are empirical questions, even if they have *obvious* answers..." (ibid.; my emphasis). In this passage what is the significance of the clause "...even if they [i.e. the questions Erwin asks] have *obvious* answers" (ibid.; my emphasis)? Erwin does not state explicitly that he believes that the questions he asks will have "obvious answers". On the other hand, it seems that a reasonable interpretation of what he has in mind is that the answers will be obvious and in the affirmative (i.e. that people *do* act on the desire to get a drink; or that a desire to get a drink sometimes *does* make a difference to the way a person behaves). If the latter is the case, then I think it shows up Erwin's scepticism to be to a significant degree a pretence. Having argued against the legitimacy of relying upon typical FP causal attributions, Erwin should not be entitled to assume that the answers to questions about the actual causal role of desires etc. will be "obvious" if, as seems, that is what he is suggesting here. Instead, Erwin's sceptical position demands that, pending the scientific demonstration of (for example) whether the desire to get a drink ever makes a difference, causally, to the way someone behaves, we should not have any settled opinion on the matter, either for or against. This is because it is a consequence of Erwin's scientific scepticism that without the scientific testing and validation of the hypotheses at issue, we simply do not know whether putative causes (such as desires) are effective in the way commonsense psychology supposes them to be.

(ii) Erwin argues that the kinds of causal connections which Hopkins assumes to be part of our everyday FP knowledge, or believes can be justified by the account of FP-reasoning he (Hopkins) provides, need to be securely validated. Erwin says: "It is one thing to make sense of what people do by postulating motives; it is another thing *to establish* that a motive featuring content overlap with an action really did cause it" (Erwin ibid.; emphasis in original). There seems to be an implication here that the putative causal connections need to be established, where this means validating the corresponding causal hypotheses to scientific standards. This in turn can only mean by appropriate experiments, since there would be no other way of testing the hypotheses to the requisite standards. It is noteworthy, however, that Erwin suggests no positive

proposals for how this might be done. Why is this? Could it be that there are currently no practical means available for testing (and hence for potentially "establish[ing]" [ibid.]) the kinds of motivational hypotheses at issue to the standards which Erwin insists need to be met, and which he criticises Hopkins for not meeting?

It would be excessive at this juncture to present a detailed analysis of the problems involved in attempting to test the putative causal relations rigorously through experiment. In chapter 5 I discussed at some length the kinds of problems that would face such testing of a category II causal hypothesis, α_1. The task of carrying out rigorous experimental tests of the kinds of hypotheses used in Hopkins's examples (e.g. the causal relevance of an individual's desire to drink a glass of water for his/her act of doing so [a category I hypothesis]) would be formidable, if not currently impossible in practice. Overall, I do not think that Erwin conveys anything like the true scale of the difficulties that would be involved in 'establishing' (i.e. validating to scientific standards) motivational hypotheses of the kind used in Hopkins's examples. Erwin himself offers no practical method for how this might be done.

7.22 ANTISCEPTICISM

A second set of attitudes constitutes a position that is, in effect, the antithesis of the first. Whereas sceptics of FP knowledge cast doubt on its genuineness or on the degree of reliability with which we possess it, antisceptics tend to conclude that it is exceedingly unlikely that we could *not* be in possession of genuine FP knowledge. Moreover, they tend to maintain that our possession of genuine and reliable FP knowledge is abundant and routinely obtained in the normal circumstances of social life. The latter is, of course, standardly maintained alongside an acceptance of the fallibility of FP claims. In addition, an antisceptic may well accept that the problem of being able to provide a fully adequate justification for FP claims is unresolved. Antisceptics include, in my judgement, Hopkins, Fodor, Grünbaum and Wellman (see e.g. Hopkins [1988] and [1991]; Fodor [1987] preface and chapter 1; Grünbaum [1993] 163-164; Wellman [1990]). Also, I include myself in this category.

By placing credence in what he has less than full justificatory warrant for it may appear that the antisceptic is behaving irrationally or is begging the question. However, I do not think that such a negative appraisal is deserved. This is because although the justification which he can provide may not be perfect or complete it may still be sufficiently strong to support his general position.

As we have seen, the main justificatory problem for the antisceptic is that, typically, he cannot validate *particular* FP claims (be they singular or generic) to standards that the sceptic seeks. For example, if a sceptic calls into question whether a given desire really is causally relevant for a subsequent action (Erwin's example) or that an individual's mental state really is about some particular intentional object and not some other (Worrall's example) then the antisceptic will be hard pressed to find a demonstration of these which is 'watertight'. The antisceptic may appeal to intuitive or informal arguments such as that, by and large, we can rely on an individual's self-proclaimed knowledge of his own mental states, or on his discernment of their causal role in actions. However, this will be seen by the sceptic as begging the question. However, in spite of the shortcomings of the antisceptic's ability securely to justify particular FP claims (and I have no suggestions of my own that could fully resolve the

sceptical uncertainties raised by Erwin and Worrall), the sceptic is far from having matters his own way. This is because there are other - more *global* - types of defence of the reliability of FP knowledge which can be appealed to. Let us consider some of these:

(a) Firstly, various philosophers have commented on how powerful and reliable FP is as a system for explaining and predicting behaviour. These comments do not answer the kinds of detailed requests for justification that sceptics such as Erwin and Worrall ask for. However, they do lend plausibility to the conclusion that, as a system of inferences, FP-reasoning must be up-and-running and working effectively (though imperfectly). The latter is a conclusion that could not have been consistently drawn from the paralysing mire of doubt about the standing of FP claims which the sceptics led us into.

Typical comments of this type - making a case for the general reliability of FP inferences and their predictive and explanatory success - are provided by, for example, Fodor:

> "Commonsense psychology works so well it disappears. It's like those mythical Rolls Royce cars whose engines are sealed when they leave the factory; only it's better because it isn't mythical. Someone I don't know phones me at my office in New York from - as it might be - Arizona. 'Would you like to lecture here next Tuesday?' are the words that he utters. 'Yes, thank you. I'll be at your airport on the 3 p.m. flight' are the words that I reply. That's *all* that happens, but it's more than enough; the rest of the burden of predicting behaviour - of bridging the gap between utterances and actions - is routinely taken up by theory. And the theory works so well that several days later (or weeks later, or months later, or years later; you can vary the example to taste) and several thousand miles away, there I am at the airport, and there he is to meet me. Or if I *don't* turn up, it's less likely that the theory has failed than that something went wrong with the airline. It's not possible to say, in quantitative terms, just how successfully commonsense psychology allows us to coordinate our behaviors. But I have the impression that we manage pretty well with one another; often rather better than we cope with less complex machines.
>
> The point - to repeat - is that the theory from which we get this extraordinary predictive power is just good old commonsense belief/desire psychology. That's what tells us, for example, how to infer people's intentions from the sounds they make (if someone utters the form of words 'I'll be at your airport on the 3 p.m. flight', then, ceteris paribus, he intends to be at your airport on the 3 p.m. flight) and how to infer people's behavior from their intentions..."
>
> (Fodor [1987] 3; emphases in original)
>
> (Note: the theory which Fodor refers to here is the *tacit* theory of

intentional mental states which, *ex hypothesi*, is utilised when we reach conclusions about others our ourselves in FP terms - Fodor op. cit. 2-3.)

And;

> "The moral so far is that the predictive adequacy of commonsense psychology is beyond rational dispute; nor is there any reason to suppose that it's obtained by cheating. If you want to know where my physical body will be next Thursday, mechanics - our best science of middle-sized objects after all, and reputed to be pretty good in its field - is *no use to you at all*. Far the best way to find out (usually, in practice, the *only* way to find out) is: *ask me!*"
> (Fodor op. cit. 6; emphases in original)

(b) Another type of argument for the overall success and reliability of FP hinges on considerations of what the likely consequences would be for social communication and learning if FP did *not* operate routinely on correctly made inferences and judgements[44]. It may be put as follows: if sound FP inferences were not routinely made, and true FP conclusions were not routinely reached, then social communication and learning, as well as interpersonal understanding, would not be possible or would break down. It is held to be a fact that human beings do genuinely communicate, understand one another psychologically and learn from one another. It is also assumed that their ability to do the latter is crucially dependent upon distinctive FP 'mechanisms' (i.e. psychological processes or states of a kind which theorists of FP postulate). Consequently, it must be the case that FP inferences are largely sound and FP conclusions are largely true. In other words, if we accept that there is genuine communication or social learning, and that this involves FP mechanisms, then it is very likely that many of the FP hypotheses which are structural constituents of that communication or learning are true. A similar verdict for many of the simple hypotheses of (say) emotional or motivational understanding (by means of FP) would then seem to be a natural extension of the above. If a sceptic objects to the claim that there is substantial genuine communication etc. between human beings then we should ask him whether he deems his present effort successful. If he answers affirmatively then he supports our claim in this instance; if he answers negatively then we should ask him why he replies as he does since, presumably, his response is predicated on his having understood us.

The foregoing argument can be deepened in an interesting way. The deeper variant first of all points out that all advanced or sophisticated forms of intellectual endeavour (such as science and the philosophy of science) are based upon the assumption that - routinely and reliably - we understand one another's mental states (such as beliefs and desires). Without such understanding social learning would not be possible, and without social learning we could not learn science, philosophy, scientific method etc.. Therefore, if the assumption that social learning is crucially dependent upon FP competence is correct, if one has learned science and its methodology then one must have already made extensive - and reliable - use of FP. Hence, according to this argument, if one purports to be intellectually competent in any science or in the methodology of science, it follows that one will already have made an extensive and reliable use of FP inferences and judgements. This vindicates the general effectiveness and reliability of FP and its reasoning.

Someone could, of course, claim that he/she acquired competence in science or its philosophy *without* a social process of communication and learning, but this is so implausible as to be not worth taking seriously (even Albert Einstein needed to learn mathematics and physics, and there are no known asocial means of doing this). Alternatively, it could be objected that whereas social communication and learning are necessary for attaining competence in science etc., such communication and learning are not predicated upon FP skills, as was assumed in the foregoing argument. This is a potentially more serious objection but, assuming that we are not ontological sceptics of FP, it is unlikely to be warranted. It is implausible to suggest that interpersonal communication and understanding are not predicated upon FP competence: for example, in ascertaining what another person believes, or in correlating that person's beliefs with his/her intentions or actions (see below). If a critic denies this then he will need to argue for it.

The argument in (b), above, can be illustrated and supported with an example - a young pupil learning mathematics (many other examples could have been used). My main points are, once again, that: (i) attainment in any intellectual field requires social communication and learning; (ii) such communication and learning is

dependent upon the successful and reliable employment of FP skills; consequently, (iii) if one has attained competence in any intellectual field (such as science) this strongly suggests that one has already been a competent user of FP, thereby vindicating the reliability of FP inferences, judgements and conclusions on a routine basis. The example is as follows.

Jimmy is seven years old and is learning mathematics in a classroom. Jimmy's ability to learn that (say) 5 x 7 = 35 is dependent upon his ability to ascertain reliably the intentional object of his teacher's belief (cf. Worrall's sceptical objection) and also, having done that, that what the teacher believes is causally relevant for her action of what she writes on the blackboard (cf. Erwin's sceptical objection). Suppose that Jimmy was a Worrallite sceptic. Then, rather than infer that what the teacher's belief was *about* when she said that "5 x 7 = 35" was (that) 5 x 7 = 35, he might well conclude that (when she said "5 x 7 = 35") what her belief was actually about was (say) 5 x 7 = 34, or even 3 x 4 = 12. Indeed, given Worrall's objection, it is even possible that when the teacher says "5 x 7 = 35" the intentional object of her belief is that 'My train was late this morning'! Being able routinely and reliably to infer what the intentional objects of other people's beliefs are is, I maintain, in general an *essential* requirement for learning. (This is not, of course, incompatible with the claim that there are kinds of learning that do not require making inferences about the intentional contents of other persons' mental states.) As stated earlier, I accept that Worrall is making an intellectually valid point about the fallibility of FP knowledge-claims, as well as about the problem of adequately demonstrating what is claimed: there is a genuine epistemological difficulty here. On the other hand, I maintain that Worrall cannot be correct if he thinks that this (admitted) inadequacy of our justificatory capabilities implies that we do not, *as a matter of fact*, regularly and routinely acquire correct knowledge of other people's mental states and of what those mental states are about. That there could be such an enormous gap between what (it seems) we can routinely and reliably know and what we can adequately provide satisfactory justification for is, I admit, a puzzle. However, I think that this is precisely the way things are when it comes to a great many judgements and conclusions of FP - we *do* know or infer them correctly, yet we are (currently) incapable of being able to provide 'watertight'

justifications for them.

A comparable response can be made to the type of sceptical objection which Erwin raises. Erwin draws attention to the problem of how can we know whether an FP desire is causally relevant for a subsequent action, even when the desire is "acted on intentionally" (Erwin [1993] 446). He rejects reliance upon informal judgements and intuitions of commonsense in such cases and implies that we need to "establish" the putative causal connection (Erwin ibid.).

Consider now what would follow if Jimmy could not - routinely and reliably - infer the causal relevance of his teacher's mental states for her actions. (Although I have no adequate account of *how* Jimmy can do this I believe that he *does* do it, routinely and reliably. If he could not he would not be able to learn elementary mathematics; yet he does do that.) After the teacher says "5 x 7 = 35" she says "I'm now going to write that down on the blackboard". She then commences to write 5 x 7 = 35 in chalk on the blackboard. If Jimmy could not reliably infer that the teacher's intention (her mental state of intending) - and, more specifically, her intention to write what she means and believes by the utterance "5 x 7 = 35" - was causally relevant for her subsequent action of writing 5 x 7 = 35 on the blackboard, then why should Jimmy make any connection between what he ascertains is his teacher's belief (i.e. that 5 x 7 = 35) and the marks she draws on the blackboard (i.e. a '5' followed by an 'x' sign, followed by a '7' etc.)? Perhaps her writing the mark '5' followed by an 'x' etc. (i.e. 5 x 7 = 35) on the blackboard was an action for which a quite *different* belief and intention were causally relevant (e.g. the belief that 3 x 4 = 12, and the intention to write *that* on the blackboard [assume that the teacher had, five minutes earlier, expressed *this* belief and also her intention to write what she meant by *it* on the blackboard]). My overall point is that unless Jimmy possesses a highly sophisticated and largely successful inferential capacity for ascertaining the causal relevance of mental states for actions (*pace* Erwin), it becomes altogether mysterious how Jimmy could possibly relate the mental states which he infers his teacher to have (i.e. her beliefs, desires, intentions etc.) to her specific actions. Yet without the ability to relate correctly such mental states to specific actions (in which the former are judged to be causally relevant or irrelevant for the latter) the patterns of inference necessary for learning would break down.

7.3 GRÜNBAUM'S ACCOUNT OF FP CAUSAL RELEVANCE EVALUATION FOR THE CASE OF INSULTS

Grünbaum is by no means opposed to the idea that genuine FP causal knowledge can be acquired by inference in natural (i.e. 'real-life' social) settings. In his brief comments on FP (e.g. Grünbaum [1984] 206, 221, 229; [1993] 112, 163-164) there is no indication of the ontological scepticism towards FP expressed by, for example, Paul Churchland (Churchland [1988] 43-49, [1989] chapter 1), or the epistemological scepticism which surfaces in at least one place in the writings of Edward Erwin (Erwin [1993] 446). Also, when commenting on a paper by Nisbett and Wilson ([1977]) in which one of the authors' theses is that human beings make FP causal attributions not by direct introspection but by inference, Grünbaum says:

> "Nor...[are Nisbett and Wilson]...driven to claim that the stated common sense verdicts on the cognitive causes of the specified affective states are generally false or shallow."
>
> (Grünbaum [1980] 360) (N.B. Grünbaum had been considering examples such as: attributing sudden experienced grief to learning that a loved one had died; or attributing experienced elation to learning that one has won a lottery.)

Grünbaum's intention in this passage is not primarily to provide an endorsement of the reliability of FP causal conclusions but, instead, of the principal role played by inference (as opposed to introspection) in reaching them. Nevertheless, his implication that "common sense verdicts" on the causes of mental states need *not* be construed as "false or shallow" (ibid.) provides evidence of his willingness to construe those (FP) inferences as sound and theoretically significant. When taken in combination with his other, admittedly brief, comments (Grünbaum: references given above) we are, I believe, justified in concluding that Grünbaum has a moderately favourable view of the scope and reliability of FP inferential competence for at least some categories of causal FP hypotheses. This in turn, I believe, justifies our classifying him as an antisceptic.

In this section I shall focus on a single passage from Grünbaum's [1993]. The passage is important because in it Grünbaum puts forward a view as to how he believes that at least one category of causal FP hypotheses can be validated. The example he uses is the causal relevance of derogatory remarks (i.e. insults) for one's introspectively experienced emotions (such as anger). Grünbaum says:

> "...I can validate by means of [Mill's] methods a supposed causal link between my conscious awareness of a derogatory remark and my introspectively experienced anger by noting, over a period of time and repeatedly, that the presence vs. the absence of insults *makes a difference* statistically to my becoming angry."
> (Grünbaum [1993] 164; emphasis in original)

This is what I shall refer to as the 'key passage'. Grünbaum immediately goes on to say (in the next sentence):

> "Similarly, by finding that people who do not drink coffee recover from the common cold no less than coffee drinkers do, Mill's methods can be used eliminatively against the causal hypothesis that observed coffee drinking for, say, two weeks cures the observable symptoms of a cold."
> (Grünbaum op.cit.)

In this section I shall argue that Grünbaum's account in the 'key passage' is unlikely to be correct. I have no objection in principle to the idea that Millian or Millian-like logic could be involved in FP causal reasoning in some cases or at some stages of such reasoning. However, I think that any conclusion that it (or any other specific strategy of causal relevance inference) is, as a matter of fact, used must be guided by empirical considerations and must not be made simply by an *a priori* decree. Grünbaum does not provide any discriminating evidence in support of his suggestion that the method he describes is, as a matter of fact, how human individuals infer the causal relevance of insults for emotions in real-life settings. His proposal that it is by the use of Mill's methods seems to be based solely on a methodologist's appreciation that Mill's method's are a way of inferring causes. But there is a genuine empirical (or naturalistic) question about how human cognition is operating so as to be able to make the appropriate causal relevance discriminations in a social environment. We are not concerned *just* with the methodologist's abstract logical question about how causal

relevance relations can be inferred (though we *are* concerned with this too). Rather, we are faced also with a question about human cognition: namely, how does the causal reasoning which is used (and this may be Millian or of some other type) fit into the psychology? FP-reasoning does need to be considered in terms of its abstract inductive-logical capabilities. This enables us to maintain a methodological perspective on it as a set of inferential capabilities. At the same time, these capabilities seem to be domain-specific to a significant degree. For example, FP-reasoning seems to be specialised for dealing with inductive inferences about (primarily) intentional mental states and, when these are the mental states of others, there may be specialisation for being able to reach conclusions about those states in particular practical situations (those found in the social environment). If this is correct, these specialised contexts in which FP-reasoning operates need to be understood. They also need to be related to the wider biological perspective in which FP-reasoning capabilities are viewed as evolved traits. Attempts to understand - or to model - FP-reasoning must therefore operate simultaneously on two fronts: one which theorises about the implicit logic; and the other which takes into account all relevant empirical (e.g. cognitive and social) and practical facts about its operation. Overall, my view is that any reconstruction of FP causal reasoning which is likely to be adequate needs to satisfy two broad criteria:

> (i) It must provide an account of the logical principles by means of which the making of causal relevance inferences is possible. (This is a methodologically normative requirement).
>
> (ii) It must fit the facts of the way human cognition functions in practice. This implies how it functions in real-life social environments. (This is a requirement for empirical or naturalistic accuracy).

In my opinion, Grünbaum's account of FP-reasoning is unlikely to be satisfactory by the second of these criteria. However, because reconstructions of FP-reasoning are likely to require the best integrated solution to both (i) and (ii), inadequacy with respect to cognitive or practical facts may require revision of the

account of the tacit logic employed. Consequently, if Grünbaum's account of FP-reasoning in the case of insults is inadequate from an empirical point of view, this may require revision of the specific account of the causal-inference logic that he suggests it involves.

My main point, then, is that Grünbaum's account is very likely to be incorrect because it is very likely to be descriptively inaccurate.

Firstly, I take it that when Grünbaum says (in the key passage) that this is how he can validate a supposed causal link between an insult and his subsequent emotion(s) that he literally means that that is how he (or any normal person) can do so (or that this is how, in normal circumstances, we routinely do so). If Grünbaum were not making such a factual, descriptive claim (in addition to a methodologically normative one) then his assertion would amount to little, and could certainly be ignored by anyone concerned with the problem of how, as a matter of fact, we do make FP causal relevance evaluations in practice in 'real-life'.

But if Grünbaum is making a factual claim in the key passage then, I maintain, the evidence strongly suggests that it is false. According to Grünbaum's model an individual is capable of evaluating the causal relevance of insults for his subsequent emotion(s) by means of Mill's methods – presumably by the Method of Agreement or possibly by some combination of the Methods of Agreement and Difference. But, if this were the case – if Mill's methods really were being used – the individual would have to "[note] over a period of time and repeatedly" whether "the presence vs. the absence of insults makes a difference statistically to [his/her] becoming angry [or experiencing some other emotion]" (Grünbaum ibid.).

Now, is it factually accurate that when one is insulted – and, let us assume, one subsequently experiences anger or embarrassment – one evaluates the causal relevance of the insult for one's emotion in the way Grünbaum describes? My central objection is that if Grünbaum's account is correct it would take a *considerable number* of occasions of being insulted in order to be able to ascertain whether or not the insults were causally relevant for the particular type of emotion under consideration (e.g. anger). Because Grünbaum's model requires several or many cases of being insulted in order to make the requisite causal relevance discrimination I shall label it the

'Multiple Instance Model' (MIM). Using Mill's methods, one would need to compare the presence vs. the absence of insult-events (say, insult speech acts; or the uttered insult-expressions one hears) with the presence vs. the absence of particular emotion-experiences (i.e. occurrences of feeling anger, embarrassment, or whatever). I do not know precisely what sample size (i.e. number of occasions of being insulted) would be needed to make a reliable causal relevance evaluation on the basis of Grünbaum's model. However, let us assume it was twenty (this may be a conservative estimate given the other factors which, conceivably, might be causally relevant for one's emotions). Then, on the basis of that model, it would take twenty occasions of being insulted before (to use myself as an example) I can begin reliably to infer whether or not, on the first occasion, the insult was/was not causally relevant for any emotion that I subsequently experienced. Moreover, even after those twenty occasions all I get is a generic result which informs me (generically) of the causal relevance of insults for a particular type of experienced emotion (e.g. anger) within the compass of the sample used. This would still leave me with the problem of how I could use that generic knowledge to infer whether or not on, say, the first occasion of my being insulted, I should have rationally attributed causal responsibility for my emotion(s) on that occasion to the insulter.

In my judgement (based on my personal experience of insult-phenomena) the above scenario is unrealistic. It seems to me that once our cognitive faculties have been 'tuned' to be receptive, we are standardly able to ascertain the causal relevance of an insult for our subsequent emotions fairly reliably even after a single insult (or a very few insults). Because Grünbaum's account describes comparing the presence vs. the absence of (respectively) insult-occurrences and emotion-occurrences "over a period of time and repeatedly" (ibid.) it would seem that it implies a much lengthier process.

7.4 THE SINGLE-CASE MODEL FOR INSULTS

According to my Single-Case Model (SCM), it ought standardly to be possible for an insultee to evaluate the causal relevance of an insult for his subsequent emotions even after a single insult (once his cognitive faculties have been 'tuned' to be receptive). In making this claim I make some assumptions: for example, that the insultee is linguistically competent and well-versed in the culture within the framework of which the insult is made; that the insultee is reasonably emotionally sensitive yet has sufficient intelligence to consider such obvious alternative hypotheses as that he might have seriously misheard or misunderstood what was said or intended. Also, I am assuming that the insultee is rational, emotionally balanced and of good will (and not, for example, paranoid or seeking to pick a quarrel at the least excuse). Equally, I am restricting myself to straightforward cases of insults in which the insult was intended, unambiguous and spoken clearly, and in which it has the potential to offend if understood.

Since I am claiming that a reliable evaluation of causal relevance is possible in the single case (even though, of course, any such undertaking is fallible), it is a burden of my thesis to show how this could be methodologically possible without vitiation from *post hoc ergo propter hoc* or other fallacies of causal inference. I believe that I have a plausible argument for showing how that latter is possible, although it is not my claim that no generalisations (say, in the form of tacit background knowledge) are involved.

My model makes crucial use of several assumptions:

(i) Insults are essentially meaningful phenomena: that is, they are linguistically information-bearing;

(ii) Insults are, possibly, causally efficacious (for having certain effects on victims);

(iii) The linguistic meaning of the insult is a central feature of its

(possible) causal efficacy;

(iv) Emotions are (in the cases I shall be considering) *intentional* phenomena – that is, they have an intrinsic cognitive component such that they can be *about* specific states-of-affairs designatable by a linguistic expression, s.

Whereas Grünbaum may well take (i) to be obvious I do not know what his views are on (iii) and (iv) (he does not discuss them). (Also, he obviously holds (ii) to be true.) However, what is pertinent for our purposes is that in his Multiple Instance Model neither (i) nor (iii) nor (iv) are made to do any *work*. Grünbaum does not single out the linguistic meaning of the insult as having any special role in terms of possible causal efficacy, and he appears to treat emotions non-intentionally simply as perceptible qualitative states.

In contrast, I think that it is a psychological fact that many (perhaps the vast majority) of occurrences of emotional states are intentional, and that this is a feature of their semi-cognitive nature. A philosopher who defends an intentionalistic view of the emotions is Robert Gordon ([1987]), and I am broadly in agreement with such a position (see pages 227-228 of this thesis).

Regarding assumption (iii), this can be thought of as a theoretical postulate for which there is some evidence. It is a widespread commonsense belief that the meaning of an insult is central to its causal efficacy. This, indeed, is a factor exploited by those persons who are unkind enough to devise specific insults calculated to hurt or humiliate in specific ways. One can also carry out a thought-experiment to appreciate the likely truth of (iii). Imagine two separate insult occurrences directed at some unfortunate individual. The first is a normal insult in which the insult-utterance, p, has a hurtful or humiliating meaning or connotation. Now consider a second situation in which an actor stands in for the insulter and in which the accompanying bodily behaviour and situational factors are exactly as before but, on this occasion, the insulter says something wholly innocuous (e.g. p = "the capital city of France is Paris!"). Although the actor's behaviour and the situational factors may contribute to the

emergence of some distress in the insultee, we would expect that he/she would not be offended as much as if a linguistically significant insult had been voiced. This, I think, supports the commonsense view which is articulated in assumption (iii).

The methodological possibility that was claimed above (i.e. reliable causal relevance evaluation in the single case) is predicated upon matters being as described in points (i)-(iv), and exploits those circumstances in a crucial way. Essentially, the basic ideas of my model are as follows:

If someone insults me and if there is a subsequent change in my emotional state (as compared to the prior 'resting' emotional state), it is very likely that the newly arisen emotion will have an intentional structure (I consider only cases in which it does). There is an important role played here by the self-monitoring of one's mental states. If there is no change from a 'resting state' then the question of seeking a cause (for the absence of a change) does not normally arise in practice.

If a new emotion arises then it ought to be possible through introspection to specify, at least roughly, what the intentional object of that emotional state is (this will be representable by a linguistic expression, s). Suppose, for example, that Sam insults me by saying "You dress like a scarecrow" (we shall represent the insult expression by 'p'). Suppose also that a moment after hearing the insult I experience (newly arisen) embarrassment. Will the embarrassment be just a qualitative feeling? My claim is that, typically, the answer will be "no". Typically, I will feel embarrassment at or about something fairly specific. In the present case let us suppose that the intentional object of my embarrassment is (roughly speaking) my somewhat scruffy attire (assuming that I am dressed in such a way). In such a circumstance note that s = p', where p' is thematically very similar to the insult p. p states that "[Dash] dress[es] like a scarecrow"; p' is equivalent to or implies that "[Dash] is scruffily dressed". Note also that psychologically, in this example, p' involves a self-conscious recognition of the possible truth of p, allowing for idiomatic comparisons – "scarecrow" meaning "scruffily dressed [person]".

The above analysis has been made for embarrassment. We may need somewhat different analyses for different emotions. For example, if my emotional

response was one of anger, then it is unlikely that the intentional structure would be of exactly the same form as for embarrassment. If Sam insults me with the expression p then (for psychological reasons which we do not need to consider here) it is likely that my emotion of anger will be directed at Sam for having said p (or for having harboured the thought that p) rather than being directed at a state which is characterisable by p'. That is, in the case of embarrassment the form (i.e. the intentional structure of the emotion) is:

'[Dash's] embarrassment at p''

(Where p' = '[Dash's being] scruffily dressed')

Whereas in the case of anger the form is:

'[Dash's] anger at [Sam's] saying or harbouring the thought that p'

(Where p = '[Dash] dress[es] like a scarecrow')

In spite of such individual variations for different emotions the basic methodological principle of my model (see below) is the same in each case. In each case there is a comparison made between the content of the insult-expression (p) and the content of what one's emotion is about (s). This is even if - as was the case with anger - there is an embedding of the significant content, p, in a secondary structure at which the emotion is directed (i.e. in the case of anger, s = '[Person X's] saying or harbouring the thought that p'). For simplicity's sake I shall continue to use the example of embarrassment in what follows.

My methodological argument is based on the following consideration: how likely is it that an emotion about (specifically) p' (= s) will have arisen in me shortly after my having heard the insult p (directed at me) and it *not* be the case that the insult-utterance is positively causally relevant for my emotion? My claim is that, given background knowledge and assumptions such as those articulated earlier, it is very

unlikely (i.e. very likely that the insult *is* causally relevant for the emotion). In order to appreciate this it is necessary to realise that after having heard the insult, if a new intentionally-structured emotion arises in me, the intentional object, s, of that emotion could have taken any of an indefinite number of linguistic expressions as its 'value'. For example, after Sam insults me by saying "You dress like a scarecrow" I might have felt embarrassment, but it could have been embarrassment at (say) "[Dash] is frightened of flying" or "[Dash] is slow at learning" etc.. The fact (if it is a fact) that a particular emotional response with a definite intentional object s arises in a person after a given insult-event is a *contingent* matter. Since it is contingent, the content of s can, I claim, serve as a form of evidence in making a probability judgement about the likely causal relevance of the insult (which has content p) upon it. Of course, the above example was for illustrative purposes only. In a real-life situation we would have to wait to see what emotional response (if any) followed upon a particular insult-event (for a given insultee), and what the content of that emotion is (if there is one).

It is assumed that 'in real-life' the inference regarding the likely causal relevance of the insult for the content of one's emotion (if one responds emotionally) will be conducted as an automatic cognitive operation. The inference will utilise the following principle: if there is a close thematic correspondence between the content of the insult (p) and the content of one's emotional response (s) then the probability of the insult being positively causally relevant for the latter is increased; and conversely if there is little or no thematic correspondence.

To avoid possible misunderstanding the following should be borne in mind: it is not thematic correspondence *alone* that justifies the conclusion of a likely positive causal relevance relation. There are a whole range of tacit or explicit background considerations which influence the making of the inference and the degree of its reliability. For example (as indicated), there is a background assumption that, logically, any one of an indefinite number of possible linguistic expressions could have been represented as the intentional content of a newly arisen emotion. My claim is that the thematic correspondence between the content of the insult and the (intentional) content of any ensuing emotion is *amongst* the factors (even though I think it is a centrally important one) that can be used in making a probabilistic inference regarding

the causal relevance of the insult for one's subsequent emotion. This point is worth emphasising because Grünbaum has drawn attention to a fallacy of causal reasoning (the Thematic Affinity Fallacy – see e.g. Grünbaum [1984] 55, 227-228; [1993] 138-139) to which, I maintain, the strategy used in my model is not susceptible. According to that fallacy:

> "...it is always fallacious to infer a causal linkage between thematically kindred events from their *mere* thematic kinship."
> (Grünbaum [1993] 138; emphasis in original)

However, as I have attempted to show, in my model a complex set of inferences is being utilised. This includes use of thematic correspondence, but there is no suggestion that that is all that is involved, or that thematic correspondence would be sufficient for inferring a causal connection.

7.5 POSTSCRIPT

Following a seminar presentation of the ideas in 7.3 and 7.4 (L.S.E. February 1997) Nancy Cartwright and Gabriel Segal raised a problem for my Single-Case Model. Firstly, they emphasised that when thinking about emotions we need to take into consideration the qualitative experiential aspect (i.e. the feeling of 'what it is like' to be experiencing the emotion). This, they pointed out, is distinguishable from the primarily cognitive – or attitudinal – component which had played the central role in my analysis. Thus, according to Cartwright and Segal, a comprehensive analysis of emotions would require us to take cognisance of both: (i) the experiential component (or, colloquially, the 'feely' part); and (ii) the propositional attitudinal component.

With the foregoing distinction in mind, Cartwright and Segal argued that whereas my model might provide some *prima facie* ground for believing that it would be possible to infer the causal relevance of the insult-utterance for the attitudinal component in the single case, it does not do so for the feely component. Their reason is connected with the apparent autonomy (or detachability) of the two components. Attitudinal and feely components can be thought of as essentially separate on this construal – analysis of the causal relevance of a third variable (i.e. the insult event) for one of these says nothing about its possible causal relevance for the other. Suppose a matching of content between p (the insult-statement) and s (the intentional content of the attitudinal component of the emotion) can provide a means of evaluating the causal relevance of the insult for the attitudinal component. Even so, according to the criticism, this inferential strategy does not extend to cover the possible causal relevance of the insult for the feely component (which, of course, lacks any such attributable propositional content). Unless it can be shown that the insult-utterance is causally relevant for the feely component there is no reason to suppose that there is a causal relation between them (it is, after all, possible that the feely component was caused by something other than the insult-utterance). According to Cartwright and Segal, my model does not provide any basis for evaluating the causal relevance of the insult for the feely component. Also, a defender of Grünbaum's Multiple Instance Model might argue that that model aimed at providing an account of how insult-occurrences can be

inferred as being causally relevant for, specifically, the feely component of emotions.

I accept that Cartwright and Segal have raised an important objection, and one that leads to further issues about the nature of emotions as well as to how, in everyday life, we infer the causal influence (or lack of such influence) of external factors upon them. In what follows I do not purport to be able to provide a fully satisfactory response to their objection. Instead, the following discussion is intended only as a provisional stand-in for such a reply, which is beyond the scope of what can be undertaken in the remaining space. The divisibility of an emotion (or particular emotion-occurrence) into experiential and attitudinal components is a deep and consequential one and I suspect that neither Grünbaum's account nor my own (for the case of insults) will prove to be entirely adequate.

There is one implication of Cartwright and Segal's criticism which I think is undoubtedly correct: namely, that emotional phenomena are more complex than I had done full justice to in my analysis. In my model, based on that analysis, I deliberately gave prominence to the cognitive (and attitudinal) aspects of emotions. This was because I had felt that these had been neglected by Grünbaum, and yet could make a difference to how causal relevance evaluations could be carried out 'in real-life'. I think it is also true that in my analysis I did not make sufficiently prominent the extent to which emotional phenomena are experiential (as opposed to attitudinal), even though I had, of course, been aware of their experiential nature. To give due attention to the experiential aspect of emotions might require recasting the whole problem of how external factors (such as insults) can be informally inferred as being causally relevant for emotions. This is not to say that I think I was wrong about a possible role played by the correspondence of content (or lack of it). However, I think that Cartwright and Segal are right that that factor, by itself, cannot be taken as evidence for the causal relevance of the insult-utterance for the feely part of an emotional response, especially if the latter is conceived in its most 'visceral' sense (e.g. the feeling of 'one's blood boiling with anger').

If we consider the experiential component of emotion alone, then the potential augmentation of the inductive inference through the matching of content clearly cannot apply, since qualitative experience has no such (semantic) content. In this

case it does seem that, in order for the individual to be able to learn through his own experience about the possibility of a causal connection between insult-occurrences and emotion-occurrences he must at some stage undergo a process in which the presence vs. the absence of the former is juxtaposed against the presence vs. the absence of the latter. This must occur either repeatedly, or in some other way such that the influence of extraneous variables can be discounted.

From this does it follow that Grünbaum's account in the key passage was correct after all, and that his Multiple Instance Model is vindicated? I would not go so far as to admit this.

Firstly, Grünbaum is less specific than I have been about what exactly is involved in either an insult or an emotion. I singled out the meaning of the insult-utterance for special attention, and as conceivably having a distinctive role to play, causally, in generating an emotional response (distinct, that is, from other aspects of the insult phenomenon such as behavioural or situational factors). Grünbaum does not do this. In his account there is no analysis of how what the insult means could evoke a specific emotional response which is related to the meaning. Similarly, Grünbaum does not treat emotions intentionally, but appears to treat them simply as qualitative perceptual states. As a general point (and not as a specific criticism of Grünbaum), I think that any approach which does the latter will sooner or later come to be seen as inadequate. This is because (in my judgement) the nature of human emotional experience is such that it most often occurs interfused with cognition, so that the emotions are intentionally directed (whether or not one is immediately aware of the intentional object). An account of emotions which views them as basically intentional is given by Gordon. He says:

> "A good place to begin the investigation [of emotions] is to note a feature that nearly all so-called emotions share with beliefs and desires: Instances of an emotion such as anger are understood to *have reference to the world*. Suppose someone tells us that Mary is very angry. Or *something* tells us: the eyes, the mouth, the content of her conversation, or the vocal inflections. Immediately we want to know: "Whom with?" and "What about?" The more fundamental question seems to be 'What about?' For if we knew what she was angry *about*, it would generally be right to assume that she is angry *with* the person or persons - or, more broadly, agent or agents - she believes *responsible* for whatever it is she is angry about.
>
> Why is it so important to answer the question 'What is she angry

> about?' For one thing, it appears that no one is truly angry unless he or she is angry about something, however general that something may be. So, finding a plausible answer to this question gives us more confidence in our initial hypothesis, that Mary is angry. More important, we need an answer if we are to anticipate Mary's actions. Without at least an inkling of what she is angry about, we can anticipate only the general accompaniments of anger, or more particularly the demeanor that generally accompanies anger *in Mary*.... With no idea what she is angry about, we can only wait and see what her anger will motivate her to do.... Much the same can be said for most of the other so-called emotions. Unless we have some idea what a person is afraid *of*, we gain little of predictive value in learning that a person "is afraid." Moreover, if our concern is to see to it that there are or are not recurrences of a certain type of emotion - in others or in ourselves - we had better establish what the emotion is about."
>
> (Gordon [1987] 22-23; emphases in original)

Secondly, there is a question about whether Grünbaum's account in the 'key passage' is naturalistically accurate even if we restrict ourselves to the experiential component. Grünbaum speaks about how his ability to validate the (possible) causal connection involves noting his "conscious awareness of [the insult]" and his "introspectively experienced [emotion]" (ibid.). However, I do not think that it can be taken for granted that the relevant causal relevance evaluation is undertaken in the way described. Grünbaum's account gives the impression that the evaluation can be carried out as a conscious and even deliberative activity. It is at best uncertain that this is how human beings can (and do) make causal relevance evaluations when their being insulted is followed by subjective emotional experience. We cannot rule out that various psychological mechanisms are involved that would require a modification of Grünbaum's account. For example, it is possible that various types of behavioural conditioning effects or cognitive operations provide inferential shortcuts to having to conduct the inference in the manner described by Grünbaum (the same also applies to my own model). Also, the causal relevance evaluation may be at least in part subconscious, as opposed to wholly conscious. (These latter suggestions are, of course, speculative and may turn out not to be correct.)

In spite of the limitations of the models discussed, it is essential that models (or theories) of how FP-reasoning operates, and of how FP claims could possibly be tested and justified by FP-reasoning, continue to be proposed. These models should be influenced (and, as far as possible, tested) by empirical considerations. However, as indicated, they should also incorporate plausible accounts of the tacit logic

or methodology utilised in FP-reasoning. Without such models we would have no way of progressing in our understanding of the nature of FP-reasoning or its logical mode of operation.

CHAPTER 8

CONCLUSION

8 CONCLUSION

The issues raised in this thesis have largely concerned the relative inductive merits and demerits of canonical scientific experimentation and FP-reasoning. Although the target of critics such as Grünbaum, Eysenck and Erwin has been FPA, and not FP or FP-reasoning *per se*, the combination of their advocacy of experimentation and neglect of making a positive case for FP and FP-reasoning has resulted, in effect, in the espousal of experiment-led psychiatric methodology. Although I am wholly in favour of experiments in psychology and psychiatry whenever they can be beneficially and effectively applied, the drawbacks of an approach overwhelmingly dominated by experimentation need to be exposed.

Firstly, if such a position is prepared to grant no significant autonomy to FP-reasoning and seeks to extend psychological knowledge solely through the application of experimental methodology, then it ignores the knowledge that, discounting sceptical objections, FP-reasoning is already capable of providing. This includes a modest, though significant, amount of knowledge that can be acquired in GPC sessions. Moreover, FP-reasoning and the tacit theory of FP seem to be able to provide an understanding of intentional mental states. This includes the causal role of those mental states in actions, and their character (specified in terms of content) as being the effects of certain environmental causes. There seem currently to be no autonomous experimental methods (and no current scientific theory) that can reproduce, let alone improve upon, this capacity.

Secondly, the experimentalism of, at least, Grünbaum and Eysenck is inadequately self-critical, giving an impression of *scientistic* bias. It really is not good enough to be less epistemologically critical of experimental methodology than of psychoanalytic inference. As we have seen (section 4.4), Grünbaum and Eysenck make little or no attempt to expose various (generic) limitations and liabilities of experimental methodology which are deserving of attention. This is in contrast to the zeal with which they (very often justifiably) examine and expose the weaknesses of Freudian inference. Methodological criticism should apply even-handedly, across the board, since without that we will not know where we stand, epistemically. The experimental methods of mature scientific practice should not be exempt from a searching critical analysis.

In addition to the above, doubts about the practical applicability of Grünbaum's model of testing for causal relevance (i.e. the SMCR) have been raised (chapter 5). A further point of emphasis has been the importance of knowledge of the individual in psychiatry (chapter 6). As against the view of some scientific methodologists (see Cartwright's criticism on page 176 and in footnote 39), it has been argued that knowledge about individual persons and individual happenings can be of perfectly legitimate scientific concern. If current methods fail effectively to test hypotheses about individuals or individual occurrences, blame may have to be pinned on the methods, rather than seeking to excuse the failure by pleading that the targets of interest were scientifically illicit.

The view has been defended that FP-reasoning is in *some* respects inductively more capable than current experimental methods and than any projected improvements in such methods in the foreseeable future (section 2.1). If correct, this is an important conclusion because it dethrones experimentation from its pretence of being at the pinnacle of relevant methodologies in psychology and psychiatry. My own view is that a more accurate portrayal of the epistemological significance of experimentation and FP-reasoning requires us to put them much more on an equal footing – side by side rather than in a hierarchy in which experimentation is on top.

In spite of the advocacy of anti-scepticism with regard to FP knowledge (section 7.22) the problem of validating FP claims to demanding standards remains. The conclusions of FP-reasoning cannot, in general, stand as scientifically validated statements. This, by itself, places severe limitations on the role which the autonomous use of FP-reasoning could play in the development of psychology as a science. At the same time, FP-reasoning seems to provide us with genuine knowledge in excess of that which we can rigorously justify. This seemingly curious situation suggests that our actual inductive capabilities (at least, in the domain of commonsense psychological understanding) are a few steps ahead of our ability rigorously to justify what we can know in a variety of circumstances. This picture resonates with certain remarks made by Gilbert Ryle in his [1949] 'The Concept of Mind'. Ryle says:

> "...there are many classes of performances [i.e. acts or utterances] in which intelligence is displayed, but the rules or criteria of which are unformulated....

> ...Efficient practice precedes the theory of it; methodologies presuppose the application of the methods, of the critical investigation of which they are the products....
>
> ...It is therefore possible for people intelligently to perform some sorts of operations when they are not yet able to consider any propositions enjoining how they should be performed. Some intelligent performances are not controlled by any anterior acknowledgments of the principles applied in them."
>
> (Ryle op. cit. 30)

Overall, the role that FP and FP-reasoning ought to play in scientific psychiatry is problematic and unresolved. On the one hand, if they can yield genuine knowledge (e.g. about intentional mental states and their causal roles) then it seems illicit to debar them unless or until scientific psychology and its methods reach a stage at which they can reproduce these competencies. On the other hand, the testing and validation of hypotheses by FP-reasoning fails in general to satisfy the standards of inductive demonstration required by science. Experimental methodology cannot be regarded as wholly competent for the domain of psychology if it is ineffective for or excludes an important sub-domain, many hypotheses of which seem capable of being reasonably well tested and validated informally (i.e. through FP-reasoning). But neither can those informally validated conclusions form part of science, as traditionally conceived.

One way of viewing FP and FP-reasoning are as, respectively, proto-science and proto-scientific method (John Worrall has espoused this – personal communication). However, there are problems in characterising the relationship between FP and its reasoning, on the one hand, and developed science and its methods (in psychology) in this way. What exactly is implied by labelling FP and FP-reasoning as 'protoscientific'?

Is protoscience at one end of a continuum with mature science and its methods? If so, it is not at all evident that FP and FP-reasoning fit this model. It is far from obvious that FP could ever be developed into a scientific psychology (see e.g. Botterill and Carruthers [1999] chapter 2), or that the tacit inferential capacities of FP-reasoning, if made explicit, would be isomorphic with canonical experimental strategies, or could be made so. At the very least these connections, if they exist, need to be demonstrated.

On a different understanding of the term 'protoscience' it might be conceived as an essentially rational set of operations, but simpler (less inductively powerful or theoretically sophisticated) than mature science. In this case

protoscience has no pretensions of ever developing into something mature, methodologically or theoretically – it remains perpetually 'immature'. But if FP and FP-reasoning are to be characterised as protoscience in this sense, then there is still a problem. This is because, as perpetually immature, their relation to *bona fide* science and its methods remains unarticulated and problematic. The classical view of *bona fide* science is that it is historically progressive, involving the testing of novel theories through a rational method, and leading to ever greater explanatory success. Even though this picture of the nature of science has come under attack ever since Thomas Kuhn's iconoclastic 'The Structure of Scientific Revolutions' (Kuhn [1970]) it is, in my view (and in spite of various remaining difficulties), in important respects still preferable to its rivals (see e.g. Kitcher [1993] chapter 4). But if FP is very largely frozen, theoretically, and its reasoning (though rationally based) is largely tacit and unamenable to improvements, then it seems that they are very different from science, in this progressive sense. Moreover, the 'perpetually immature' view of FP and its reasoning does not sit comfortably with the fact – if, as I have claimed, it is a fact – that they are in some respects inductively and theoretically more competent than current scientific psychology and its methods. Protoscience which is in some respects superior to the best current developed science (and applied scientific methods) gives the impression of the word being used as an oxymoron.

As matters stand neither experimentation nor FP-reasoning can take over the epistemological function of the other. Since both, I maintain, are needed to maximise understanding and knowledge-growth in psychology and psychiatry, it would seem that methodological pluralism can be expected in these fields for the foreseeable future. That is, primary reliance will have to be placed either on experimentation or on FP-reasoning, depending on the hypothesis we wish to have tested.

Finally, although the main body of the thesis concluded with an unresolved problem (section 7.5), far from regarding this as an embarrassment, I have felt it to be apt. It stands as a metaphor for the fact that all research is open-ended. At best we resolve certain problems only to become aware of further ones.

NOTES

NOTES

1. I devised this scheme in response to a criticism by Guy Longworth of my paper 'Science and Folk Psychology in Psychiatry' delivered at the University of London Graduate One-Day Conference (22nd May 1998). Longworth's complaint was that my paper often did not make clear precisely which class of psychological statements was being referred to (in a given passage) and that, consequently, the precise nature of my disagreement with Grünbaum over the appropriate source of validation for those statements was obscured. I trust that the above classification scheme will substantially clear up this difficulty. I should also like to acknowledge here my gratitude to Longworth for his helpful comments.

2. Epidemiology is standardly taken to be that branch of medicine which examines how the frequency of diseases is distributed in the population or is related to environmental factors (Backett and Robinson [1992] 183). Epidemiological methods will include, for example, statistical analyses of the frequency of a variable within different subgroups of a population.

3. Freud seems implicitly to recognise a distinction of this kind when he says:

> "We have a right, or rather a duty, to carry on our research without consideration of any immediate beneficial effect. In the end – we cannot tell where or when – every little fragment of knowledge will be transformed into power, and into therapeutic power as well. Even if psychoanalysis showed itself as [therapeutically] unsuccessful...it would still remain completely justified as an irreplaceable instrument of scientific research."
> (Freud [1917](a) 294-295)

   There are, of course, important ethical issues raised by the question of whether psychotherapists should be allowed to practice without there being clear evidence that their activities are beneficial.

4. It is, of course, possible to frame a hypothesis about therapeutic benefit and to attempt to test this in the psychotherapy session. This, however, would simply be an instance of making epistemological usage of the session.

5. Grünbaum explains what he takes the "major" or "cardinal" Freudian hypotheses to be as follows:

> "We are told [by Freud] that repression plays the crucial *causal* roles of producing neuroses, engendering dreams, and generating a very important subclass of slips.... Thus, when I speak hereafter of the "major" [or "cardinal"] hypotheses of psychoanalytic theory in the foundational sense, I mean this cardinal trio of causal postulates."
> (Grünbaum [1993] 110-111; emphasis in original; reference omitted)

6. The example of an exception which Erwin provides comes from Behaviouristic psychology, not from psychoanalysis or psychotherapy. (It involves inferring the causal efficacy of electric shocks in treating the self-injurious behaviour of an autistic boy.) Consequently, in opposing "the sceptic" (ibid.) who denies the possibility of uncontrolled causal clinical confirmations, Erwin still concedes little if anything to the ability of psychoanalytic or psychotherapeutic case-studies to test/confirm (non-trivial) causal hypotheses. For further discussion of this example see Note 7. Also, there is at least one passage in which Erwin expresses what appears to be a sceptical attitude towards the acquisition of FP knowledge (Erwin [1993] 446). As a result of this I have described him as a sceptic of such knowledge in 7.21, where I discuss the relevant passage.

7. Erwin explicitly affirms that some clinical causal hypotheses can be confirmed in uncontrolled case-studies (and suggests that Grünbaum might agree) - Erwin [1981] 559. The examples that Erwin provides in order to vindicate this apparent concession to the competency of uncontrolled case-studies come, however, from Behaviouristic psychology (i.e. behaviour therapy) (Erwin op. cit. and [1988] 206-207). In one case mild electric shocks were used to eliminate the self-injurious behaviour of an autistic boy (op. cit.). Grünbaum uses this same example when arguing that some single-subject (i.e. case-study) causal inferences can be valid (Grünbaum [1984] 259-260; [1993] 230-231). The problem with this example is that the causal hypothesis it involves (i.e. that the electric shocks were causally responsible for the change in behaviour) is so unlike the kinds of causal hypotheses informally tested by psychotherapists (which are, for example, often mentalistic and/or postulate as causes events from the client's ordinary social environment), that it lends, in effect, no support to a

more favourable verdict on the effective testing of causal hypotheses in *psychotherapeutic* case-studies.

8. These categories are not strictly demarcated but overlap.

9. Although Edelson combines this negative appraisal of experimentation with a positive one for the psychoanalytical case-study testing of at least some Freudian hypotheses (e.g. Edelson [1984] chapter 11), I am not here endorsing the latter. My present aim is to draw attention to the problems and limitations of experimentation. If experiments cannot test (many or most) Freudian hypotheses in a highly critical and effective way it does not, of course, follow that the case-study method will be able to either.

10. I think that there are many issues concealed in the argument which would need to be carefully analysed and critically evaluated before a reliable verdict could be reached. To do this would take us too far afield, especially since it is not essential for our present needs. However, we can mention some of the issues that would need to be considered. How, precisely, is the concept of the 'strength' of a theory or hypothesis to be explicated? How is that concept related to empirical content and to confirmability or falsifiability? Is there any regularity or pattern to the way in which the 'strength' or the empirical content of a theory relates to its experimental testability, given the latter is always dependent on the practical performability (or otherwise) of the experiment, and not merely on its inherent logic (see subsection 1.4(B))?

11. For example, when in order to satisfy methodological orthodoxy an experimental study is carried out on hypothesis h' (because h' can be tested experimentally), whereas the hypothesis we would really like to have tested is h, but an effective experimental test of h could not be performed for theoretical or practical reasons. Here h' superficially resembles h but is actually substantially different from it. We met this kind of situation earlier when discussing the experimental testing of FPA claims.

12. This is not, of course, the same error as believing that experiments will always, or will necessarily, be effective in testing the hypotheses they were intended to test, or that they will always or necessarily yield reliable conclusions. This (extremely naïve) error amounts roughly to a belief in the infallibility of experiments.

13. See also Erwin [1988] 205-206 for some comments relevant to the general point.

14. Kitcher has said:

> "Placing the knowing subject firmly back into the discussion of epistemological problems seems to me to be the hallmark of naturalistic epistemology."
>
> (Kitcher [1993] 9)

15. It is not being suggested that FP-reasoning is a completely 'rigid' or 'mechanical' natural inductive process. There does seem to be some capacity for the exertion of autonomous (or conscious) analytical judgement or reasoning over the domain it covers. However, this is limited and would seem to be possible only with sufficient cognitive maturity (for issues of relevant interest see Gopnik [1996] 178-180).

16. If the conclusion reached is a product of one or more natural inductive processes plus conscious (or autonomous) analytical reasoning then, insofar as that latter reasoning can also err, there will be added scope for error.

17. It could even be suggested as further hypotheses that you were surprised and/or puzzled to find this printed sentence occurring here. Are these further hypotheses true or false? Surprise (though possibly not puzzlement) has affective colouration. Hence, if you were surprised and I was able to infer this it follows that by means of FP-reasoning I would have been able to learn something about your emotional (and not just your cognitive) state.

18. In h_1 the emotion is fear or anxiety; and the intentional content, which is a feature of the accompanying cognitive character of the emotion – i.e. in this

case, what is feared – is "being left all alone in the world". For each of the other hypotheses h_2-h_6 the respective emotions and their intentional contents can easily be inferred.

19. The justification for a fairly good level of reliability for many FP- and GPC-inferences is provided at various places in the thesis (see especially sections 2.2, 2.3, 2.4, chapter 3, chapter 7). It is not being suggested that this amounts to a complete or watertight justification (one which successfully rebuts all critical objections). The latter probably cannot be provided. What is offered are various arguments which attempt to show that many of the conclusions which are arrived at by FP- (or informal) inference are deserving of rational credence, and that any systematic scepticism is undeserved.

20. There may be some degree of prejudice and intellectual double standards in this attitude. Background epistemic assumptions to experimental investigations are also extensive and yet may not have received any independent justification. It cannot simply be taken for granted that they will be potentially less problematic than those involved in a question and answer procedure of the above kind.

21. Unless there is some other effective modality for testing them apart from FP-reasoning or experimentation.

22. In a written comment to a draft version of this section (Autumn 1998) John Worrall asked what I would think about a lie-detector test. There are at least two problems with this:

    (i) Lie-detector tests are, of course, fallible;

    (ii) What lie-detector tests seek to establish is whether a person is lying. But lying (or telling the truth) is a specific kind of cognitive operation – and one that is different from the task of independently testing whether the proposition under consideration (i.e. about which the person is either lying or telling the truth) is itself true or false. The problem I posed was not: "Are there ways of testing whether I am lying or not when I claim (as I have) that p is true?" It was, rather, a more fundamental one: "Are there any independent experimental means of testing whether p is true or false?" At its best, a lie-detector test could reliably

establish whether I am telling the truth (or lying) when I say that p is true/false, given my prior cognition of it as true or false. Presumably, in practice, such tests require a preliminary 'calibration'. The calibration involves careful measurement of the individual's physiological or behavioural responses (e.g. heart-rate, electrical conductivity at skin surface etc.) in response to certain calibrating questions. These will be questions the truth or falsehood of the answers to which are already known: e.g. "Is your name [so and so]?"; "Do you have a degree in law"?; "Have you a sister named Mary?" etc.. Certain assumptions will be made (e.g. that the test-subject clearly understands the questions, and answers truthfully). We can accept these assumptions for present purposes.

Given the above, it should be noted that using my response to infer the truth/falsity of the hypothesis (in this case p) is already predicated upon my ability to know whether the hypothesis is true or false (through subjective recognition). The test does not seem to have the ability directly to discern the truth or falsehood of p by measuring physiological indicators, scanning my brain, or whatever. (It is not clear that there is any constant correlation between the truth-conditions of a proposition – be it about external circumstances or mental contents – and, say, changes in heart-rate, brain activity etc..) Instead, as far as establishing the truth/falsity of p is concerned, the lie-detector test is epistemically parasitic upon my capacity for introspective knowledge. Consequently, it does not serve as an autonomous scientific (experimental) test of p, one which by-passes reliance on my subjective recognition or knowledge.

The important point is that it appears that there is (currently) no way of by-passing epistemic reliance upon the test-subject's own capacity to know whether p is true or not in certain cases. These will, typically, be cases in which: (i) the individual concerned was the only witness to the external event (and there is no other way to test for the event's occurrence – this also applies to '(ii)'); or (ii) the individual was the only truthful and willing witness to the external event (because other witnesses either lie or refuse to comment); or (iii) the 'event' in question is not an external, public one, but is part of the individual's perception or thought (so that the subject alone has direct access to it).

Even if, as I have claimed, in these cases the hypothesis that the event in question took place cannot be independently tested by experiment, it does not follow that it did not take place – and neither does it follow that the individual

who was witness to the event cannot *know* (reliably) that it did. On the other hand, an individual could certainly be mistaken in remembering, even if his aim was virtuous and he sought to disclose the truth. If there was no independent test of the claim we would not be able to utilise that means to find out the truth. But, as I have indicated elsewhere in this thesis, a wholesale rejection of reliance upon individual testimony (given the lack of independent – and especially scientific – testing) is epistemically flawed (see pages 84-85 and section 2.4).

23. The psychotherapy sessions will, however, be conducted within a value-system (typically, one of liberal humanism). Thus, although no moral judgement will be passed on involuntary psychological productions, or on instinctual impulses which are just that, the client is treated as a moral agent. The latter implies that he/she is capable of making choices, and is treated as being to a greater or lesser degree aware of a domain of ethical principles to which human conduct is referred.

24. I think it is beyond doubt that psychoanalysts and other psychotherapists have sometimes been guilty of this kind of fault. Freud seems to be guilty of it in, for example, his 'Dora' case (see Eysenck [1985] 62-65). However, from this it does not follow that there is no capacity for the undistorted assimilation of data in psychotherapy sessions and judicious theoretical interpretation of that data. Indeed, there may even be a substantial capacity for the latter.

25. In spite of this, Tom's subjective assessment was that the limited number of sessions he had were of considerable benefit to him and constituted a 'turning point' for the better in the progress of his disorder.

26. There may be scientifically conducted studies of the psychological effects of aggressive or derisory behaviour on 'victims'. If so, and if the studies are methodologically sound, it would be essential to take account of the results. The point here is that even in the absence of any such studies there are very often compelling reasons to regard the results of FP-reasoning on the matter (using FP evidential sources) as broadly reliable.

27. A more thorough account would reveal that there are very interesting problems which arise in connection with analysing the causal relevance of external factors on emotional states when those emotional states are construed as having both: (i) a qualitative experiential (or 'feely') component; and (ii) a cognitive (or intentional) component. These problems are discussed in section 7.5. For present purposes, and at the present level of analysis, Hypothesis T2 is simply intended to capture the idea that if Dan (or his accomplices) had not raised, specifically, the theme that they did (i.e. about Tom's father's invalidity) during the playground incident it is extremely unlikely that Tom would have felt humiliation or anxiety about specifically that theme at that time.

28. It was, of course, standard for Tom's father not to run errands because of his incapacitation. It was Tom's mother who would be collecting him.

29. This episode is interesting in its own right from the point of view of descriptive psychology. This is because if Tom's account is reliable it shows that even at the age of just over seven years at least some children are capable of having an acute awareness of societal attitudes, including those that they perceive as unfavourable towards themselves or towards individuals or groups with whom they perceive themselves as being identified. Tom said that, in retrospect, he had no reason to suppose that Bill's mother held any negative or hostile attitudes towards disabled people.

30. Davidson says:

> "What is the relation between a reason and an action when the reason explains the action by giving the agent's reason for doing what he did? We may call such explanations *rationalizations*, and say that the reason *rationalizes* the action.
>
> In this paper I want to defend the ancient – and commonsense – position that rationalization is a species of causal explanation."
>
> (Davidson [1963], reproduced in Davidson [1980] 3; emphases in original)

31. It might be felt that to accept FP motivational explanations as a species of causal explanation would be the natural (and expected) attitude of someone who, like Grünbaum, has a strongly pro-scientific attitude. This is because, since natural science often seeks causal explanations, a common ground might be felt to exist

between it and the construal of motivational explanation in causal terms. Also, construing motives as causal might be felt implicitly to oppose any anti-naturalistic tendencies inherent in the tradition which opposes reasons to causes. However, there are also sufficient differences between action explanations (in terms of motives) and the kinds of explanation that are provided by natural science, as traditionally conceived, to render any automatic assimilation of the former into the latter questionable. For example, action explanations not only explain causally (if we accept a Davidsonian position), but are also capable of providing intelligibility, in the sense of showing why it was reasonable or rational for the agent to perform the action (Moya [1990] 106-107, 112-113). Consequently, I do not think that it was a foregone conclusion that Grünbaum *would* accept FP motivational explanations as able to supply the genuine causes of actions.

32. From this it can be seen that I have endorsed Millian strategies for inferring causal relevance in this example of an FP generalisation (see also page 38). In light of the fact that I elsewhere criticise one or other aspect of Grünbaum's espousal of (broadly) Millian strategies as operative in FP causal reasoning (sections 7.3, 7.4 and 7.5), some defence needs to be provided against any potential charge of inconsistency.

    Firstly, I *accept* that Millian strategies provide a plausible model for many cases of FP causal inferences at a fairly crude level of 'first approximation'. It is when we begin to take into consideration some of the finer details of real-life examples (perhaps especially in the single case, or when the content of intentional mental states is involved) that complications and difficulties emerge. These problems tend to suggest that a more sophisticated model (or models) of causal inference are required, and that Millian strategies , strictly conceived, are unlikely to be wholly adequate. The above endorsement of Millian strategies should therefore be viewed only in this 'first approximation' sense, and as glossing over various difficulties which are discussed elsewhere.

33. It should be noted that whereas this is the mainstream view there are some dissenters. Peter Urbach, for example, has argued not merely that retrospective clinical trials (using historical controls) are epistemically legitimate, but that

they can have advantages over (prospective) randomized clinical trials in some respects (Urbach [1993] 1429-1431; Howson and Urbach [1989] 153).

34. It is true that the dependent and independent 'variables' in each of T1-T4A are not as precisely defined (or do not correspond to such precisely described physical events) as is possible for some causal hypotheses. They are not atypical, however, of many commonly occurring variables in the psychological and social domains. I have assumed that the above features do not vitiate the legitimacy of the causal status of the hypotheses concerned.

35. I am very grateful to Colin Howson for bringing home to me the limitations of relying upon p-values for representing the outcomes of clinical trials (see e.g. Freeman [1993]). This is just one example of the need for major improvements in the way statistical analyses are made and presented in the clinical field.

36. One way of viewing this is as a logical extension (and extreme case) of the subgroup problem (mentioned in (c), above) in which the subgroup contains just one individual. However, one's conception of an individual is not, of course, limited to thinking of him/her simply as an element of a statistical sample.

37. Psychoanalytic theory postulates the causes of such slips as being *repressed* (Grünbaum [1984] 202-205).

38. For discussion of this problem see e.g. Cartwright [1989] chapter 1.

39. The full text of Cartwright's communication is as follows:

"There is though one major standard objection that I think would be raised: and I think you will either have to answer it or reorganise your thinking in some way to deal with it. The objection is: what you are doing here is accusing scientific method of being incapable of doing something it was never meant to do in the first place: to provide a test for the correct explanation of individual happenings. Science, after all, so the line goes, is meant to establish general claims, not claims about individuals. And the standard statistical methodology is,

indeed, "doing this". This point of view goes back at least [to the] Methodenstreit. Could history or political economy be a science? The argument was that it could not be, because it studies individual happenings in all their individual peculiarity, whereas genuine science aims to establish universal connections between repeatable features (or perhaps it's better to think of them as recurring features).

More recently, in the literature on causality, the distinction has been clearly marked by a number of authors: I talk about generic vs. singular causal claims; Elliott Sober and Ellery Eells talk about type vs. token claims; Paul Humphreys takes up the distinction; etc. etc.

Now there was a period when the Methodenstreit re-emerged in Anglophone philosophy, that people like Hempel insisted that science could indeed explain individual happenings by use of the deductive [nomological] method. Nobody at that time, in that debate, really took the problem of overdetermination seriously. I mean that the problem that arises when an individual falls under two laws, either of which is sufficient to determine it to have the given behaviour. So long as purely universal laws were at stake, the problem did not seem to matter much. I do not think that Hempel ever had much of a view about what to do about this problem when one was using not D-N, but rather I-S explanations. At any rate, the kinds of causal explanations you are looking at are now widely thought to be statistical and everyone is aware of the problem.

One of the people who first made it salient in this round of literature was Wesley Salmon with his example of the man who took birth control pills and didn't get pregnant. I guess you also know that this [is] the central theme of our Modelling in Physics and Economics project: we are concerned about how one uses scientific laws to provide accounts of singular cases, but nobody thinks that the methodology which is good for testing the law-like claims will simply double as a good methodology for determining the correctness of the singular claims.

Now this isn't to say that you do not have a valid point that standard scientific methodology doesn't do nearly so much for us as we would like. It's just that there is already a conventional way of looking at this problem. That seems to...put you in the position of either recasting what you are doing in this

more conventional frame, or explaining why you think that is not the right way to look at it.

I have only one other major remark. As I am sure you realise, it doesn't follow from the fact that the standard methodology for testing law-like claims does not double as a methodology for testing singular claims, that the singular claims will best be established by informal or folk methods. Again, I think our modelling project is relevant here. What we've been arguing is that there seems to be no mechanical procedure for modelling in the individual case. We also have been subscribing to the fashionable doctrine that local knowledge of a lot of concrete facts and local regularities will make a big difference. I am not at all sure that it's useful to try to divide this knowledge into scientific and non-scientific, nor to claim that the method itself is either scientific, or not scientific. For instance, one of your cases seems to use the same methodology that I ascribe to the gyroscope experiment in *Nature's Capacities*: one good way to establish that a particular causal factor is the right explanation for a given behaviour in a given individual, is to eliminate all the other possible causal factors that might have been operating in that case. The standard scientific method that you discuss comes in establishing what are the set of possible causal factors for that kind of behaviour. The method is supposed to establish, at a generic level, what kinds of things can cause what other kinds of things. Now I am sure that your intuition that we have a lot of informal information that we can bring to bear on a given case, helps us to figure out what kinds of causal factors might have been operating, is very hard to recast in the form of generic causal laws without looking all together trivial. I, myself, have not figured out any way to cash in on that intuition yet.

About this issue, do you know the views of John Sutton? His work criticises the standard economics approach which he calls the "crossed-sticks approach"; conventional economics tries to write down a system of linear equations – in this case he is thinking of two – whose simultaneous solution will give us a single point. That point is supposed to be the behaviour which is observed to occur. Sutton, instead, aims at best to give sets of axioms that constrain the solution; they map out only regions in which the actual behaviour must fall. What determines which point will actually occur, are other kinds of historical factors, claims Sutton. Of course, he doesn't have a well worked out methodology for

how these historical factors enter and play their role, or how we are to test specific hypotheses about given occurrences – which is your problem, after all. But I thought reminding you of this point of view might be of some help."
(Dictated by Professor Cartwright, 21 December 1995; minor typographical errors in the original have been corrected in the above.)

It should be borne in mind that Cartwright's comments relate to an *earlier* version of chapter 5 in which I had taken a stronger view of the apparent failure of canonical scientific (and especially experimental) methods to test hypotheses such as α_1 effectively. In the earlier version I had, for example, said:

"If there was a failure of [testing for an] hypothesis it would be easy to jump to the conclusion that the hypothesis must be to 'blame' and that the deficiency could not be on the part of mature science or its methods. Indeed, in the confident mid-Twentieth-Century heyday of science and its philosophy this might have been the natural inference to make. However, such an optimistic (not to mention arrogant) view of mature science and its capabilities is, I urge, unwarranted. The criteria which give meaningfulness and empirical content to a hypothesis do not automatically guarantee that it will be testable by the methods of mature science. Here we have been considering whether $\{\alpha_1, \alpha_2, \alpha_3 \ldots \alpha_n\}$ might constitute such a class of empirically sound hypotheses, untestable not merely by [applications of] the SMCR [i.e. Grünbaum's 'Standard Model of Testing for Causal Relevance'], but by any [mature scientific method (MSM)]. The belief that if such untestability arises it could not possibly be mature science or its methods which are deficient or limited seems to stem from an entirely bogus view of the latter: for example, that extant MSM's have limitless testing powers, or that the only hypotheses worthy of our interest are those which MSM's can test. There would appear to be no rational basis for either of these suppositions.

Overall, the view advanced here is that we should be prepared to view not merely the theories of mature science but also its methods of testing or validation as capable of failure – possibly on a massive scale.

A fall-back position which might be appealed to by those who have a sanguine attitude about the power of science (and, for some, its almost limitless power) is to argue along the following lines (here I am modifying the argument to fit the case for mature scientific *methods* rather than its usual form which is for the *explanatory* capacity of mature science):

(A) Even though (let us assume) we *currently* do not have any MSM capable of adequately testing some class of hypotheses Ψ, that is no argument against us not being able to devise some *future* MSM which will be adequate for the task. This claim may have appended to it a conditional claim to the effect that:

(B) If there exists (or will exist at some future time) a method which is capable of adequately testing Ψ, then that will be a scientific method.

The first part of this claim (i.e. (A)) is simply promissory or hopeful. What it hopes for – namely, the future invention or discovery of an adequate MSM – may or may not come to pass. It certainly provides no *reason* for believing in the future success of scientific methodology for the task at hand and may deflect us from thinking more deeply about methodological limitations of established scientific practice.

The second part provides a "no lose" situation for the concept of science, and is worse than useless as a means of enlightening us about the nature of science, its methods or its rationality. It is worse than useless in this respect because:

(i) It provides no advance characterisation or definition of what is to count as science (or as 'scientific') – at least, not in any sense which characterises that aspect of the putative future method which is both novel and 'scientific'; and,

(ii) It is nevertheless prepared to assimilate whatever new method(s) which may arise in the future and is adequate for the task at hand as being 'scientific', thereby making the concept of scientificity used merely definitional *post hoc*.

With such a strategy 'science' cannot lose (or cannot lose much). But the penalty for relying upon such a strategy is a lack of real penetration into understanding the nature of science and its methods and how these are related to knowledge and rationality. It also involves complacency about how scientific methods may fail. (B) acts, in effect, as a 'placeholder' for accruing whatever is epistemically favourable, worthy or productive under the heading 'science'. It will be a general background assumption of this thesis that any rational

conception of science and its methods ought to recognise them as having limitations as well as powers."

(From the first draft of chapter 5 of this thesis; Autumn 1995)

40. I have not assumed that Cartwright herself subscribes to the views she presents as an objection. This is because of the considerable importance she attaches to the singular and particular (even within a scientific framework of thinking) in her own philosophical writings – see e.g. Cartwright [1983] 19; [1989] 2-3, 9, chapter 3). Also, as can be seen from the full text of her personal communication (see Note 39), it is clear that Cartwright believes that some kind of rational or quasi-scientific explanation of singular cases is possible. She says: "...[in our Modelling in Physics and Economics project] we are concerned about how one uses scientific laws to provide accounts of singular cases..." (ibid.).

41. That is, for descriptions to be scientifically significant they would have to bear resemblance to other descriptions: *totally* unique descriptions of phenomena (if possible at all) would have no cognitive significance to us as bases for conducting investigations.

42. In the view of some philosophers the singular occurrence of X and of Y does not debar them from being causally related. It is just that if X, Y were completely novel, we could not *know* if they were causally related or not – an epistemological limitation. On this view causes are conceived as relations holding between single (unique) events in which there is an ontological dependence of the later event upon the earlier, such that it would not have occurred had the earlier event not occurred. This is clearly very different from the Humean tradition, according to which constant conjunction (implying repeated occurrences of X and of Y) defines the causal relation. Examples of philosophers who admit singular causation are Cartwright ([1989] 2-3, chapter 3) and Humphreys ([1989] 12-13, 99).

43. Of course, the search for scientific laws is not a metaphysical activity – laws of nature only receive scientific credibility insofar as there is empirical evidence for them. But the belief that the world is so structured that there exist currently

unknown underlying laws which are capable of successfully explaining 'surface' or phenomenal properties does involve metaphysical assumptions.

44. The specific arguments in this subsection are, to the best of my knowledge, my own. They are, however, based on my personal reflections on the general idea that knowledge – including scientific knowledge – is in some significant way grounded in social relations. I have never studied this position in detail (I believe it is central to the views of, for example, the later Wittgenstein). Personally, I would never advocate this position as the best way to found epistemology. This is because I think that individual human cognition in relation to the world is more important for epistemology than, specifically, social relations. Nevertheless, I think that there is an important element of truth in the view that social relations influence the knowledge that we possess or are capable of possessing.

REFERENCES

REFERENCES

ALDERSON, M. [1983] *An Introduction to Epidemiology*. Second edition. London: Macmillan.

ALLPORT, G. W. [1961] *Pattern and Growth in Personality*. New York: Holt, Rinehart and Winston.

ALLPORT, G. W. [1962] 'The General and the Unique in Psychological Science'. *Journal of Personality* 30: 405-422.

ALVAREZ, W. [1997] *T.rex and the Crater of Doom*. Princeton, New Jersey: Princeton University Press.

BACKETT, S. and ROBINSON, A. [1992] 'Epidemiological Methods'. In C. Freeman and P. Tyrer (eds.) *Research Methods in Psychiatry: A Beginner's Guide* [1992] pp.183-207.

BARON-COHEN, S. [1995] *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge (Massachusetts) and London: MIT Press.

BARON-COHEN, S. and SWETTENHAM, J. [1996] 'The Relationship Between SAM and ToMM: Two Hypotheses'. In P. Carruthers and P. K. Smith (eds.) *Theories of theories of Mind* [1996] pp.158-168.

BERNSTEIN, D. A., ROY, E. J., SRULL, T. K. and WICKENS, C. D. [1991] *Psychology* (2nd edition). Houghton Mifflin.

BOLTON, D. and HILL, J. [1996] *Mind, Meaning and Mental Disorder: The Nature of Causal Explanation in Psychology and Psychiatry*. Oxford and New York: Oxford University Press.

BOTTERILL, G. and CARRUTHERS, P. [1999] *The Philosophy of Psychology*. Cambridge: Cambridge University Press.

CAMPBELL, D. T. and STANLEY, J. C. [1963] *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally and Company

CARTWRIGHT, N. [1983] *How the Laws of Physics Lie*. Oxford: Clarendon Press.

CARTWRIGHT, N. [1989] *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

CARRUTHERS, P. and SMITH, P. K. (eds.) [1996] *Theories of theories of Mind*. Cambridge: Cambridge University Press.

CHASSAN, J. B. [1967] *Research Design in Clinical Psychology and Psychiatry*. New York: Appleton-Century-Crofts.

CHASSAN, J. B. [1979] *Research Design in Clinical Psychology and Psychiatry*. Second edition. New York: Irvington Publishers, Inc..

CHURCHLAND, P. M. [1988] *Matter and Consciousness*. Revised edition. Cambridge (Massachusetts) and London: MIT Press.

CHURCHLAND, P. M. [1989] *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge (Massachusetts) and London: MIT Press.

CLARK, P. and WRIGHT, C. (eds.) [1988] *Mind, Psychoanalysis and Science*. Oxford: Basil Blackwell.

DAVIDSON, D. [1963] 'Actions, Reasons and Causes'. In D. Davidson *Essays on Actions and Events* [1980] pp.3-19.

DAVIDSON, D. [1980] *Essays on Actions and Events*. Oxford: Clarendon Press.

DAVIES, M. and STONE, T. (eds.) [1995] *Folk Psychology: The Theory of Mind Debate*. Oxford: Basil Blackwell.

DREYFUS, H. L. (ed.) with HALL, H. [1982] *Husserl, Intentionality and Cognitive Science*. Cambridge (Massachusetts) and London: MIT Press.

DENNETT, D. C. [1987] *The Intentional Stance*. MIT Press.

EARMAN, J. (ed.) [1983] *Testing Scientific Theories*. Minneapolis: University of Minnesota Press.

EARMAN, J., JANIS, A. I., MASSEY, G. J. and RESCHER, N. (eds.) [1993] *Philosophical Problems of the Internal and External Worlds*. Pittsburgh and Konstanz: University of Pittsburgh Press

EDELSON, M. [1984] *Hypothesis and Evidence in Psychoanalysis*. Chicago and London: The University of Chicago Press.

EDELSON, M. [1986] 'The Evidential Value of the Psychoanalyst's Clinical Data'. In A. Grünbaum 'Précis of *The Foundations of Psychoanalysis: A Philosophical Critique*' [1986] pp.232-234.

EDELSON, M. [1988] *Psychoanalysis: A Theory in Crisis*. Chicago and London: University of Chicago Press.

ERWIN, E. [1980] 'Psychoanalysis: How Firm is the Evidence'. *Noûs* 14: 443-456.

ERWIN, E. [1981] 'The Truth About Psychoanalysis'. *The Journal of Philosophy* 78: 549-560.

ERWIN, E. [1986] 'Psychotherapy and Freudian Psychology'. In S. Modgil and C. Modgil (eds.) *Hans Eysenck: Consensus and Controversy* [1986] pp.179-203.

ERWIN, E. [1988] 'Psychoanalysis: Clinical Evidence versus Experimental Evidence'. In P.Clark and C. Wright (eds.) *Mind, Psychoanalysis and Science* [1988] pp.205-223.

ERWIN, E. [1993] 'Philosophers on Freudianism'. In J. Earman et al *Philosophical Problems of the Internal and External Worlds* [1993] pp.409-460.

EVANS, S. J. W. [1993] Discussion of 'The Value of Randomization and Control in Clinical Trials' by P. Urbach. *Statistics in Medicine* 12: 1433-1436.

EYSENCK, H. J. [1953] *Uses and Abuses of Psychology*. London: Penguin.

EYSENCK, H. J. [1985] *Decline and Fall of the Freudian Empire*. Harmondsworth: Penguin Books Ltd..

EYSENCK, H. J. and WILSON, G. D. [1973] *The Experimental Study of Freudian Theories*. London: Methuen.

FERGUSON, R. D. G. and FERGUSON, J. G. [1994] 'Cerebral Intraarterial Fibrinolysis at the Crossroads: Is a Phase III Trial Advisable at This Time?'. *American Journal of Neuroradiology (AJNR)* 15: 1201-1216, August 1994.

FODOR, J. A. [1987] *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge (Massachusetts) and London: MIT Press.

FØLLESDAL, D. [1982] 'Brentano and Husserl on Intentional Objects and Perception'. In H. L. Dreyfus (ed.) in collaboration with H. Hall *Husserl, Intentionality and Cognitive Science* [1982] pp.31-41.

FREEMAN, C. and TYRER, P. (eds.) [1992] *Research Methods in Psychiatry: A Beginner's Guide*. Second edition. Gaskell Publishers, for the Royal College of Psychiatrists, London.

FREEMAN, P. R. [1993] 'The Role of *P*-Values in Analysing Trial Results'. In *Statistics in Medicine* 12: 1443-1452.

FREUD, S. [1909] 'Notes Upon a Case of Obsessional Neurosis'. In A. Richards (ed.) *Freud – Case Histories II* [1979] pp.36-128. The Pelican Freud Library Vol. 9. Harmondsworth: Penguin Books.

FREUD, S. [1916-17] *Introductory Lectures on Psychoanalysis*. The Pelican Freud Library Vol. 1 [1973]. Edited by J. Strachey and A. Richards. Harmondsworth: Penguin Books.

FREUD, S. [1916] 'The Premisses and Technique of Interpretation'. Lecture 6 in S. Freud *Introductory Lectures on Psychoanalysis* [1916-17] pp.129-142.

FREUD, S. [1917](a) 'Psychoanalysis and Psychiatry'. Lecture 16 in S. Freud *Introductory Lectures on Psychoanalysis* [1916-17] pp.281-295.

FREUD, S. [1917](b) 'Fixation to Traumas – the Unconscious'. Lecture 18 in S. Freud *Introductory Lectures on Psychoanalysis* [1916-17] pp.313-326.

FREUD, S. [1917](c) 'Some Thoughts on Development and Regression – Aetiology'. Lecture 22 in S. Freud *Introductory Lectures on Psychoanalysis* [1916-17] pp.383-403.

FREUD, S. [1917](d) 'The Paths to the Formation of Symptoms'. Lecture 23 in S. Freud *Introductory Lectures on Psychoanalysis* [1916-17] pp.404-424.

FREUD, S. [1924] 'A Short Account of Psychoanalysis'. In A. Dickson (ed.) *Freud - Historical and Expository Works on Psychoanalysis* [1986] pp.161-182. The Pelican Freud Library Vol. 15. Penguin Books.

FREUD, S. [1925] 'The Resistances to Psychoanalysis'. In A. Dickson (ed.) *Freud – Historical and Expository Works on Psychoanalysis* [1986] pp.263-273. The Pelican Freud Library Vol. 15. Penguin Books.

FREUD, S. [1933] *New Introductory Lectures on Psychoanalysis*. The Pelican Freud Library Vol. 2 [1973]. Edited by J. Strachey and A. Richards. Harmondsworth: Penguin Books.

FREUD, S. [1933] 'The Question of a *Weltanschauung*'. Lecture 35 in S. Freud *New Introductory Lectures on Psychoanalysis* [1933] pp.193-219.

FREUD, S. [1934] Postcard to Saul Rosenzweig. Quoted in D. W. MacKinnon and W. F. Dukes 'Repression' (p.703), in L. Postman (ed.) *Psychology in the Making* [1964], New York: Knopf.

GLYMOUR, C. [1974] 'Freud, Kepler, and the Clinical Evidence'. In R. Wollheim (ed.) *Freud: A Collection of Critical Essays* [1974] pp.285-304.

GLYMOUR, C. [1980] *Theory and Evidence*. Princeton, New Jersey: Princeton University Press.

GOODING, D., PINCH, T. and SCHAFFER, S. (eds.) [1989] *The Uses of Experiment – Studies of Experimentation in Natural Science*. Cambridge: Cambridge University Press.

GOPNIK, A. [1996] 'Theories and modules; creation myths, developmental realities, and Neurath's boat'. In P. Carruthers and P. K. Smith (eds.) *Theories of theories of Mind* [1996] pp.169-183.

GORDON, R. M. [1987] *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge: Cambridge University Press.

GOWER, B. [1997] *Scientific Method: An Historical and Philosophical Introduction*. London: Routledge.

GREENWOOD, J. D. (ed.) [1991] *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press.

GRÜNBAUM, A. [1980] 'Epistemological Liabilities of the Clinical Appraisal of Psychoanalytic Theory'. *Noûs* 14: 307-385.

GRÜNBAUM, A. [1983] 'Retrospective Versus Prospective Testing of Aetiological Hypotheses in Freudian Theory'. In J. Earman (ed.) *Testing Scientific Theories* [1983] pp.315-347.

GRÜNBAUM, A. [1984] *The Foundations of Psychoanalysis: A Philosophical Critique*. Berkeley: University of California Press.

GRÜNBAUM, A. [1986] 'Précis of *The Foundations of Psychoanalysis: A Philosophical Critique*'. (This paper includes an Open Peer Commentary and Author's Response.) *The Behavioural and Brain Sciences* 9: 217-284.

GRÜNBAUM, A. [1990] '"Meaning" Connections and Causal Connections in the Human Sciences: The Poverty of Hermeneutic Philosophy'. *Journal of the American Psychoanalytical Association* 38: 559-577.

GRÜNBAUM, A. [1993] *Validation in the Clinical Theory of Psychoanalysis. Psychological Issues* Monograph 61. Madison, Connecticut: International Universities Press, Inc..

GRÜNBAUM, A. [1994](a) 'Freud's Permanent Revolution: An Exchange'. *The New York Review of Books* Vol. 41, Number 14, August 11, 1994.

GRÜNBAUM, A. [1994](b) 'Reply to Louis Berger's Review of *Validation in the Clinical Theory of Psychoanalysis*'. *Psychoanalytic Books* Vol. 5, No. 1, pp. 154-167.

GUTTENPLAN, S. (ed.) [1994] *A Companion to the Philosophy of Mind.* Blackwell.

HACKING, I. [1983] *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge University Press.

HART, H. L. A. and HONORÉ, A. M. [1956] 'Causation in the Law'. In H. Morris (ed.) *Freedom and Responsibility: Readings in Philosophy and Law* [1961]. Stanford: Stanford University Press.

HERSEN, M. and BARLOW, D. H. [1976] *Single-case Experimental Designs: Strategies for Studying Behaviour Change.* Pergamon Press.

HEWSTONE, M. [1989] *Causal Attribution: From Cognitive Processes to Collective Beliefs.* Oxford: Basil Blackwell.

HOFFER, A. and OSMOND, H. [1961] 'Double-blind Clinical Trials'. *Journal of Neuropsychiatry* 2: 221-227.

HOLTON, G and BRUSH, S. G. [1985] *Introduction to Concepts and Theories in Physical Science.* Princeton, New Jersey: Princeton University Press.

HOPKINS, J. [1988] 'Epistemology and Depth Psychology: Critical Notes on *The Foundations of Psychoanalysis*'. In P. Clark and C. Wright (eds.) *Mind, Psychoanalysis and Science* [1988] pp.33-60.

HOPKINS, J. [1991] 'The Interpretation of Dreams'. In J. Neu (ed.) *The Cambridge Companion to Freud* [1991] pp.86-135.

HOWSON, C. and URBACH, P. [1989] *Scientific Reasoning: The Bayesian Approach.* La Salle, Illinois: Open Court.

HUMPHREYS, P. [1989] *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences.* Princeton, New Jersey: Princeton University Press.

JOHNSON, T. [1992] 'Statistical Methods and Clinical Trials'. In C. Freeman and P. Tyrer (eds.) *Research Methods in Psychiatry: A Beginner's Guide* [1992] pp.24-61.

KAZDIN, A. E. [1980] *Research Design in Clinical Psychology.* New York: Harper and Row.

KIM, J. [1996] *Philosophy of Mind.* Westview Press.

KITCHER, P. [1993] *The Advancement of Science: Science Without Legend, Objectivity Without Illusions.* Oxford University Press.

KLERMAN, G. L. [1986] 'The Scientific Tasks Confronting Psychoanalysis'. In A. Grünbaum 'Précis of *The Foundations of Psychoanalysis: A Philosophical Critique*' [1986] p.245.

KLINE, P. [1981] *Fact and Fantasy in Freudian Theory*. 2nd. Edition. Methuen.

KUHN, T. S. [1970] *The Structure of Scientific Revolutions*. Second edition. (First edition published in 1962.) Chicago & London: The University of Chicago Press.

LINDBERG, D. C. [1992] *The Beginnings of Western Science*. University of Chicago Press.

MACKIE, J. L. [1974] *The Cement of the Universe*. Oxford: Clarendon Press.

MILLER, S. [1984] *Experimental Design and Statistics*. Second edition. London and New York: Methuen.

MODGIL, S. and MODGIL, C. [1986] *Hans Eysenck: Consensus and Controversy*. London: Falmer Press.

MOYA, C. J. [1990] *The Philosophy of Action*. Polity Press and Basil Blackwell.

NAGEL, T. [1994] 'Freud's Permanent Revolution'. *The New York Review of Books* Vol. 41, No. 9, May 12, 1994.

NEU, J. (ed.) [1991] *The Cambridge Companion to Freud*. Cambridge: Cambridge University Press.

NEWTON-SMITH, W. H. [1981] *The Rationality of Science*. Boston, London and Henley: Routledge and Kegan Paul.

NISBETT, R. E. and WILSON, T. D. [1977] 'Telling More Than We Can Know: Verbal Reports On Mental Processes'. *Psychological Review* 84: 231-259.

NORMAN, D. [1985] *The Illustrated Encyclopedia of Dinosaurs*. London: Salamander Books.

PERRY, J. [1994] 'intentionality (2)'. In S. Guttenplan (ed.) *A Companion to the Philosophy of Mind* [1994] pp.386-395.

POPPER, K. [1959] *The Logic of Scientific Discovery*. London: Hutchinson.

RADHAKRISHNA, S. and SUTHERLAND, I. [1962] 'The Chance Occurrence of Substantial Initial Differences Between Groups in Studies Based on Random Allocation'. *Applied Statistics* 11: 47-54.

REICHENBACH, H. [1951] *The Rise of Scientific Philosophy*. University of California Press.

ROSENBERG, A. [1988] *Philosophy of Social Science*. Oxford: Clarendon Press.

RYLE, G. [1949] *The Concept of Mind.* London: Hutchinson

SEARLE, J. R. [1983] *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press.

SEARLE, J. R. [1994] 'intentionality (1)'. In S. Guttenplan (ed.) *A Companion to the Philosophy of Mind* [1994] pp.379-386.

SEGAL, G. [1996] 'The Modularity Theory of Mind'. In P. Carruthers and P. K. Smith (eds.) *Theories of theories of Mind* [1996] pp.141-157.

SHAPIRO, M. B. [1961] 'The Single Case in Fundamental Clinical Psychological Research'. *British Journal of Medical Psychology* 34: 255-263.

SHAPIRO, M. B. [1966] 'The Single Case in Clinical-Psychological Research'. *Journal of General Psychology* 74: 3-23.

STRAWSON, P. F. [1952] Introduction to Logical Theory. London: Methuen.

URBACH, P. [1993] 'The Value of Randomization and Control in Clinical Trials'. *Statistics in Medicine* 12: 1421-1431.

URBACH, P. [1985] 'Randomization and the Design of Experiments'. *Philosophy of Science* 52: 256-273.

WELLMAN, H. M. [1990] *The Child's Theory of Mind.* Cambridge (Massachusetts) and London: MIT Press.

WILKES, K. V. [1991] 'The Relationship Between Scientific Psychology and Common-Sense Psychology'. *Synthese* 89: 15-39.

WORRALL, J. [1989] 'Fresnel, Poisson and the White Spot: the Role of Successful Prediction in Theory-Acceptance'. In D. Gooding *et al* (eds.) *The Uses of Experiment – Studies of Experimentation in Natural Science* [1989] pp.135-157.