UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Consensus and methodology

Speekenbrink, M.

Publication date
2005
Document Version
Final published version

Link to publication

Citation for published version (APA):
Speekenbrink, M. (2005). *Consensus and methodology*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Mar 2023

Consensus and Methodology

Maarten Speekenbrink

ISBN 90-9019495-9
Printed by Printpartners Ipskamp
Cover designed by M. Speekenbrink
Typesetting by LaTeX2e
Graphs made with R and gle

Copyright © 2005 M. Speekenbrink
All rights reserved

Consensus and Methodology

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties
ingestelde commissie, in het openbaar te verdedigen
in de Aula der Universiteit
op donderdag 2 juni 2005, te 12.00 uur

door
Maarten Speekenbrink
geboren te Delft

Promotores:	Prof. Dr. J. Hoogstraten
	Prof. Dr. J.H. van Heerden
Co-promotor:	Dr. P. Koele

Faculteit der Maatschappij- en Gedragswetenschappen

Voor Annalise
en Gabriel

> ...a story has as many versions as it has readers. Everyone takes what he wants or can from it and thus changes it to his measure. Some pick out parts and reject the rest, some strain the story through their mesh of prejudice, some paint it with their own delight.
>
> John Steinbeck, *The Winter of our Discontent*

Note

While all the chapters of this thesis were written to form a coherent whole, parts of this thesis have been, or are intended to be, published separately. Chapter 3 is a largely expanded version of Speekenbrink, M. (2003). The hierarchical theory and statistical model selection. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano & J.J. Meulman (Eds.). *New developments in psychometrics.* (pp. 331-338). Tokio: Springer. It also forms the basis for Speekenbrink, M. (submitted). On the simplicity principle in statistical model selection. Chapter 6 is a translated and expanded version of Speekenbrink, M. (2004). De ongegronde eis tot consensus in de methodologie. *Nederlands Tijdschrift voor de Psychologie, 59*, 1-11. Other manuscripts are in preparation.

Dankwoord

Aan dit proefschrift heb ik met veel plezier gewerkt. Daaraan hebben veel mensen bijgedragen, die ik dan ook graag hier wil bedanken.

Allereerst mijn (co-)promotores. Jaap van Heerden, Johan Hoogstraten en Pieter Koele – ik noem jullie in alfabetische volgorde, zowel qua voor- als achternaam, maar ieder verdient een plaats vooraan – bedankt voor het vertrouwen om mij mijn eigen weg te laten gaan. Jullie waren daarbij een bron van inspiratie. Johan, je overtuiging dat alles empirisch te onderzoeken is was aanstekelijk. Jaap, je overtuiging dat je sommige problemen beter analytisch kunt oplossen ook. Pieter, je aansporing om op bepaalde momenten gewoon een keuze te maken was een uitkomst. Jullie waren een gemêleerd gezelschap, wat altijd leidde tot plezier en soms ook een prettige onenigheid in onze besprekingen. De 'triangulation' in dit proefschrift – wetenschapsfilosofie, besliskunde, en experimenteel psychologisch onderzoek – is aan jullie te danken.

Mijn kamergenoten over de jaren heen hebben er op aangename wijze voor gezorgd dat ik RSI-vrij ben gebleven. Denny Borsboom, door heldere, amusante en soms ook pittige discussies over wetenschap en haar filosofie, alsmede ook door het even moeilijk te vinden om Kriterion te verlaten. Michiel Hol, door gesprekken over muziek, adaptief testen, van alles, en nog wat. Peter van Rijn, door gesprekken over muziek, dynamische IRT, van alles, en door nuttige latex en R tips. Sanneke Schouwstra, door gesprekken over validiteit, de gang van zaken bij psychologie en meer, alsmede ook door het uitlenen van babykleren en speelgoed.

De medewerkers van de programmagroep Psychologische Methodenleer, Wulfert van den Brink, Fred Cornelisse, Conor Dolan, Marijke Engels, Ellen Hamaker, Dave Hessen, Katarina Kouwenhoven, Wim Krijnen, Don Mellenbergh, Peter Molenaar, Harrie Vorst en Jelte Wicherts, hebben voor een prettige werksfeer gezorgd. Die sfeer is in grote mate ook te danken aan Ineke van Osch. En met name wil ik ook Wulfert, Conor en Wim bedanken dat ik bij jullie binnen kon lopen met soms toch kleine, soms toch grote vragen, waarop jullie altijd bereid waren een antwoord te geven (en dat meestal dan ook deden). Verder ook alle mede-leden van de Dynamic Factors,

Dave, Denny, Ellen, Michiel en Peter, voor het fijne bandjes-gevoel. Daarnaast wil ik ook de medewerkers van onze 'zuster-programmagroep' binnen ontwikkelingspsychologie bedanken. Met name Raoul, bedankt voor je hulp bij het afleiden van lastige distributies, ook al zijn die hier niet meer terug te vinden. De leden van het IOPS wil ik bedanken voor het mij op de hoogte houden van de nieuwe ontwikkelingen in de IRT, maar vooral voor de gezelligheid tijdens en na conferentie-diners en in Osaka (Niels, zo'n Japans capsulehotel was wel een ervaring) en Sardinië (Ingmar, Reinoud en Paqui, onze villa op de heuvel was een groot succes).

Er was ook een leven naast promoveren. Ik wil iedereen bedanken die mij hieraan heeft herinnerd. Vooral Bas, Michiel, Dave, Ilko, Cheryl en Stefan, bedankt voor jullie vriendschap; ik hoop jullie nu weer wat vaker te zien. My family in Scotland and New Zealand, thanks for being there and those fantastic holidays.

Mijn ouders, Hans en Ria, bedankt voor jullie liefde en steun vanaf mijn geboorte. Hans, je was er gedurende deze periode ook bij. Ria, bedankt voor alles.

Gabriel, je liet me zien dat sommige dingen veel belangrijker zijn dan wetenschap. Bedankt voor de verwondering, de afleiding, je lach en knuffels. Je bent mijn lievelingszoon!

Annalise, you're still my big and little wonder. Thanks for your love, support, strength, patience and positivity. We've pulled through together. We don't always agree, but now you know why that is a good thing. Never stop surprising me!

Maarten
Amsterdam, 17 april 2005

Contents

1
Introduction

For similar reasons, the science of psychology is both a wonderful and frustrating endeavour. The richness and complexity of its subject matter, the human mind and behaviour, make psychology a challenging topic with relevance for a wide range of issues. But because of the richness and complexity of psychology, it is also extremely difficult to cover more than a microscopic aspect of it in a single scientific work. Coupled with a large number of productive psychological researchers, this leads to a vast body of highly specialised studies. With specialisation comes specialised language, concepts and methods, which makes communication between specialities difficult. There is little dissemination of knowledge between the many subfields of psychology. Psychology is characterised as fragmented, and even chaotic, with 'little more order than a telephone book' (Royce, 1987).

1.1 The 'crisis' in psychology

It is a widely held conviction that, in order to be called a science, a field of inquiry should be paradigmatic (Staats, 1991; Denmark & Krauss, 2005). Since Kuhn (1970), the existence of a paradigm has been the main criterion to distinguish science from 'would-be science'. A paradigm, a disciplinary matrix of ideas, theories, examples, practices and methods, is an overarching principle which coordinates the action of scientists. Since there is no agreed on paradigm in any subfield of psychology (Eysenck, 1997), psychology can only be considered a preparadigmatic discipline. As Staats (1991) puts it

> No matter how many well-conducted experiments psychology produces, no matter the refinements of methods of data production and analysis, no matter how sophisticated the specialized apparatus and theory con-

> struction, as long as psychology's products are inconsistent, unrelated and mutually discrediting, psychology will be considered a "would-be scientific discipline," a tangle of knowledge rather than a clear-cut field of science [...] (p.910)

For over decades, there have been calls to change the current standing of psychology from a disunified discipline to a unified science (De Groot, 1990a; Krantz, 1987; Royce, 1970; Staats, 1991; Sternberg & Grigorenko, 2001). Not everyone believes such a unification is possible. For instance, S. Koch (1981) argues that psychology, by its very nature, defies unification. Because most psychological events are multiply determined, ambiguous in meaning and polymorphous, to name but a few of Koch's characterisations, different psychologists will, depending on their personal history and characteristics, perceive them in different ways. This may be why the Grand Theories of Yesteryear, such as behaviourism and psychoanalysis, have failed to unify psychology; they simply weren't for everyone.

Koch's explicit pluralism is not for everyone either, though. While many accept that a unifying theory is not expected in the near future, they do not take it as a principal impossibility (e.g. Royce, 1987). In any way, unifying psychology by means of a single, paradigmatic theory, is but one avenue to counter psychology's fragmentation. Proponents of unification point to methodological differences as a major source of psychology's fragmentation (Eysenck, 1997; De Groot, 1990b; Kendler, 2002; Rychlak, 2005; Staats, 1999). Kendler (2002) points to the divide between those who view psychology as a natural, and those who view it as a human science. This divide has a long history, beginning with the *Methodenstreit* in Germany at the turn of the 19th to 20th century (see Nerlich, 2004, for an overview). On one side stood Dilthey, who pleaded for a descriptive and analytical psychology, arguing that psychology requires sympathetic understanding (*Verstehen*) on the part of the researcher, which an explanatory psychology cannot provide. On the other side stood Ebbinghaus, who advocated a radically empirical and experimental psychology. For a long time, the argument has been settled in favour of Ebbinghaus, but the recent quantitative-qualitative debate has made the issue once again a prominent one. The divide between the quantitative and qualitative camps runs very deep, brought about by a different outlook on the aims of scientific psychology. But even within the confinements of mainstream, quantitatively oriented psychology, there is methodological disagreement. Eysenck (1997), as Cronbach (1957) did long before him, points to the difference between two methods of inference: experimental and correlational. Both methods have strengths and weaknesses, and psychology would benefit from a combined approach, instead of the mutual distrust displayed by both sides. Then there is the disagreement between those advocating axiomatic and psychometric approaches to psychological measurement (Borsboom & Mellenbergh, 2004; Cliff, 1992; Michell, 2000). And what about the ever continuing debate regarding null-hypothesis significance testing? (Carver, 1978; Chow, 1998; Nickerson, 2000; Rozeboom, 1960)

If methodological differences are the source of psychology's fragmentation, unifying psychological methodology may be its remedy. Since method is such a primary factor in any science, it certainly seems plausible that consensus on the best approach to study the human mind will bring together the disparate theories under investigation.

This stance is taken by De Groot (1990b), who proposed to achieve methodological unification by organised consensus meetings. Such consensus meetings or conferences are used widely in medicine (Kalberer, 1985; Klazinga, Everdingen, & Casparie, 1989), and with success, as the sheer number of consensus conferences indicates. The objective is to bring together experts in a scientific field in order to evaluate the current evidence and to put forward a consensus statement concerning best practice. But before implementing measures to arrive at consensus, it should be clear what is to be expected from methodological consensus. Are the products arising out of a consensually based methodology better than those arising out of different methodologies? Is the epistemic standing of unanimously held convictions firmer than that of convictions which differ to those held by others? Will methodological consensus result in theoretical consensus? Questions such as these are the topic of this thesis.

1.2 Science and consensus

The role of consensus in science is not unique to issues of unification. Consensus is at the core of Western scientific thinking, which takes science to be a superior path to knowledge and truth, because it is objective, and 'the hallmark of its objectivity is the ability to coerce rational inter-subjective agreement' (Bjerring & Hooker, 1980). Hence the definition of N. Campbell (1921): 'Science is the study of those judgments concerning which universal agreement can be obtained' (p.27). While this is apparently the first statement that explicitly connects science and consensus (Ziman, 1968), ideas concerning the relation between consensus and knowledge are much older. Classically, consensus was taken to be a direct consequence of the Scientific Method, which, upon its adoption, demands unanimous assent from those who adopt it (Baigrie & Hattiangadi, 1992; Laudan, 1984). Such a firm belief in the scientific method is also found in Peirce (1878), who went so far as to propose that truth is the opinion which is fated to be ultimately agreed to by the scientific community at the limit of inquiry.

Over the years, the epistemic standing of the scientific method has been relativised. The hope of a universal scientific method has waned with the demise of Logical Positivism and Logical Empiricism. In contrast to these schools, which viewed science as a cumulative enterprise, Kuhn (1970) painted a radically different picture of science. According to Kuhn, science moves from paradigm to paradigm, from scientific revolution to scientific revolution. Each scientific revolution results in a radical break from the preceding consensus. A switch in paradigm is essentially a switch in world-view. Changes in paradigm do not result from a comparison of competing paradigms on a common scientific criterion. Since paradigms are incommensurable, there simply is no way that they can be objectively compared on a common criterion of evidential support. The fate of a paradigm is not decided by a universal scientific method, but by the assent of the scientific community. In breaking with the idea that scientific consensus is a consequence of the scientific method, Kuhn has assigned consensus a pivotal role in the dynamics of scientific knowledge. As Ziman (1968) puts it: '[Consensus] is not a subsidiary consequence of the "Scientific Method", it is the scientific method itself' (p.9). Essentially the same stance is taken by sociologists of science

such as Merton, who refer to shared norms and values to explain the occurrence of scientific consensus (Laudan, 1984). Perhaps the strongest weight is given to consensus by social constructivists such as Latour and Woolgar. While the former parties take scientific consensus to be a rational reflection of the best of current knowledge, this latter party denies that scientific theories, whether consensual or not, can be reflections of an objective reality. According to this view, scientists construct a 'reality' through social exchange and negotiation. Consensus is not a proxy for truth otherwise defined, it defines truth itself.

To be sure, there are not only proponents of the consensus ideal. Consensus has connotations of harmony and unity, but also of totalitarianism and dogmatism. For instance, Feyerabend (1975) argued that

> Unanimity of opinion may be fitting for a church, for the frightened victims of some (ancient, or modern) myth, or for the weak and willing followers of some tyrant; variety of opinion is a feature necessary for objective knowledge [...] (p.33).

In turn, Agassi (1975) warns that the desideratum of consensus turns science into a 'Latin-American dictatorship'(p.24). A more recent critic of the consensus ideal is Rescher (1993). A common thread in these critiques is the idea that variety in opinion and mutual criticism are the driving forces of scientific inference. Since consensus is a state in which such variety and criticism are absent, it must stifle progress. On the other hand, it is hard to imagine where progress leads, if not to some ultimate answer on the nature of the phenomenon under study, no matter how distant in the future. Surely, one would hope that, once given, this answer will be striking enough to demand universal consent. This, in a nutshell, is the puzzle of scientific consensus: it is at once a highly desirable, and highly undesirable state. It is desirable, because the attainment of scientific aims should entail consensus. It is undesirable, because it marks a halt to inquiry, where inquiry should usually proceed. Consensus is an end to inquiry, but there is no way of telling whether we got out prematurely, or whether we went for the whole ride.

1.3 Overview

This thesis approaches the subject of consensus and methodology from a philosophical, decision-theoretic, statistical, and empirical direction. This multi-faceted approach reveals my personal conviction that pluralism is essential to a proper understanding of any subject. While there is a change in focus between the different chapters, effort was taken to present them in a logical order.

Chapter 2 introduces some important concepts for the remainder of the thesis, and is actually more of an introduction than the present chapter. The focus is on the problem of justification, or the question how to defend particular choices for theory and method. The main problem to be dealt with is that of underdetermination, which means that there is no sufficient basis for such a choice. The classical problem of underdetermination is that empirical evidence is insufficient, since two or more incompatible theories can be equally consistent with a given body of data. From

the decision-theoretic view of scientific inference introduced in this chapter, the more general problem of axiological underdetermination is introduced. This problem arises when, for a given collection of epistemic goals, multiple theories or methods have identical utility. The chapter concludes with a theoretical investigation of the concept of consensus. As will be argued, the usual definition of consensus as uniformity in opinion is inadequate. An adequate definition of consensus should take higher-levels of belief into account, those beliefs which pertain to the beliefs of others.

Chapter 3 describes a general theory which has been proposed to explain the occurrence of consensus and dissensus in scientific matters. This theory, called the hierarchical theory of justification, is compatible with the decision-theoretic view of scientific inference introduced in Chapter 2. The hierarchical theory of justification consists of three interrelated levels at which, and by means of which, consensus is forged. The lowest level is the factual level, consisting of all statements about the world. The theory prescribes that disagreement on this level can be resolved by reaching agreement on the methodological level. Disagreement on the methodological level is in turn resolvable by reaching agreement on the axiological level, which concerns the aims of scientific inference. The hierarchical theory is applied to the field of statistical model selection, which has been chosen because, due to the quantitative nature of statistical models, it allows for a precise implementation of the proposed decision-theoretic model. While the hierarchical theory is an useful means of overarching the separate literature on statistical model selection, it ultimately fails as a normative theory. There are two reasons for this failure. The first is that theoretical considerations must guide methodological decisions, which goes against the hierarchical structure of the model. The second reason is that problems of underdetermination also arise in statistical model selection, so that consensus on the appropriate method does not forge consensus on a theoretical level.

The problem of underdetermination has moved some to argue that social factors should be taken into account when dealing with scientific theory choice. Hence, Chapter 4 addresses the relation between underdetermination and consensus. Two experiments on the influence of consensus in situations of empirical underdetermination are described. Both experiments investigate the extent to which individual hypothesis behaviour is influenced by the hypotheses of peers under differing levels of underdetermination. Social comparison theory states that in situations of uncertainty, individuals will attempt to validate their opinions by comparing them to those of others. The first experiment shows that, as expected, individuals rely on social comparison in situations of underdetermination, but not in situations of determination. In the second experiment, the effects of differing levels of underdetermination on social comparison were studied. The expectation that higher levels of underdetermination result in stronger reliance on social comparison was supported, although the evidence for belief change following social feedback was weaker than in the first experiment. Overall, the effect of empirical information appeared much larger than that of social information.

Chapter 5 describes two experiments on collaboration in multiple cue probability learning. There are two main reasons why groups may outperform individuals when it comes to making good judgements and decisions. The first is that individuals may have partly non-overlapping information, so that the group can base its collective

decision on more information than any individual alone. The second is that the collective decision may be less subject to idiosyncratic bias than individual decisions. The experiments described in this chapter are an attempt to investigate the validity of these two reasons. It builds on some of the insights of Brunswik and his followers to derive plausible models for collective decision making. A model in which members' decisions are weighted by their confidence provides an adequate description of the group decisions. Moreover, since confidence is related to past achievement, the group decision process can be viewed as a weighting-by-achievement process.

Chapter 6 offers a critical examination of the possible role of consensus in a normative methodology. Three possible roles are distinguished: consensus as a goal, as a means, and as a criterion. It is argued that consensus fulfills none of these roles adequately, and as such has no role in a normative methodology. But this is not to say that social factors have no role in science, for they certainly do. However, striving for consensus is not the way to capitalise on those social factors. There is no need to demand unanimous belief, nor to demand dissensus for that matter. What is called for is a situation of 'mutual understanding', where all individuals have accurate beliefs regarding the beliefs of others, and can use these to validate and improve upon their own belief.

Chapter 7, finally, offers a summary and further discussion of the topics raised in the preceding chapters.

1.4 Notation

I've attempted to make the notation consistent throughout this thesis. Sets are typeset in calligraphic. Vectors and matrices are typeset in bold, with matrices denoted in capitals. Some symbols that will be used are:

$\equiv$	Definition/Defined as
$\wedge$	Conjunction/And
$\vee$	Disjunction/Or
$\neg$	Negation/Not
$\Longrightarrow$	Implication/Implies
$\forall$	Universal quantifier/For all
$\exists$	Existential quantifier/There exists
$\in$	Membership/Element of
$\cap$	Intersection
$\subset$	Subset
B_i	Belief-operator/Agent i believes
$u_{\|\mathcal{G}}(\cdot)$	Utility of $\cdot$ relative to a set of goals $\mathcal{G}$
$u_A(\cdot)$	Utility of $\cdot$ for individual A

2
Problems of justification

> All our reasoning is nothing but the joining and substituting of characters, whether these characters be words or symbols or pictures... if we could find characters or signs appropriate for expressing all our thoughts as definitely and as exactly as arithmetic expresses numbers... we could in all subjects in so far as they are amenable to reasoning accomplish what is done in arithmetic and geometry. For all inquiries that depend on reasoning would be performed by the transposition of characters and by a kind of calculus.... And if someone would doubt my results, I should say to him: 'let us calculate, Sir,' and thus by taking to pen and ink, we should soon settle the question.
>
> Leibniz, *The Art of Discovery*

Such was Leibniz' dream: if scientific reasoning is adequately formalised, all scientific disagreement can be settled in a satisfactory manner. This may seem preposterous now, but for long enough it didn't. The logical positivist programme was one attempt at such a formalisation. That the logical positivists didn't realise Leibniz' dream may now seem unsurprising. Famous incompleteness-theorems, such as Gödel's (see Nagel & Newman, 1958), show that in any consistent and sufficiently strong formal system, there are true statements which are not provable. In proving the theorem, Gödel used a self-referring statement, essentially 'this formula is undecidable in formalism F'. Gödel used an ingenious trick, so-called Gödel numbering, to make such meta-statements stateable in the language of arithmetic itself. The problem that arises is somewhat akin to the classical liar paradox. Consider the truth-value of the statement 'this statement is false'. If the statement is considered true, it must be false, while if it is considered false, it must be true. A true philosophical puzzle. Tarski (1944) tackled it by arguing that 'truth' can only be defined in a meta-language. This means that no formal language may contain its own truth-predicate. So all such languages are incomplete. For methodology, being concerned with rules for scientific inference, the question of completeness raises itself in the fol-

lowing problem: can a methodology, consisting of rules of inference and justification, justify these rules?

Justifying methods of justification is a difficult task, to say the least. Infinite regress looms, for if a rule is justified by another rule, then its justification depends on the justification of this other rule. If the justification of this rule depends on another rule, then this latter rule must be justified, and so forth ad infinitum. The justificatory chain should stop at some point, but where, and more importantly, how? One solution is to stop it at the point where there is no choice between competing rules. When there is no possible disagreement to settle, there are no conflicting viewpoints which differ in their justification. In the absence of conflict, there is no practical problem of justification. Insofar as scientific inference is viewed as a decision-problem, and justification as defending a particular decision, this seems an adequate solution. A decision implies the existence of alternative courses of action, and if there is no alternative, there simply is no decision to defend.

2.1 A decision-theoretic framework

A decision-theoretic analysis takes scientific inference as instrumental action. Such a view was for instance taken by Ellis (1988), Giere (1988), Hempel (1965), Laudan (1984) and Levi (1967, 1980), although none of these authors, except maybe for Levi, take full advantage of the tools of decision analysis. The triple 'actions', 'goals' and 'consequences' provides a simple framework which is general enough to capture most of scientific reasoning. With 'general enough' I mean that, insofar as scientific reasoning is purposeful action, which it undoubtedly is, scientific behaviour can be analysed as means towards ends. But offering a decision-theoretic analysis of scientific inference is not a trivial exercise. What are for instance the relevant actions? Scientific inference is a cognitive affair, so we should be considering cognitive actions, such as believing a certain proposition. But does the belief in a proposition really result from a decision to believe that proposition? This issue should be addressed when using decision-analysis in order to describe actual scientific inference. However, this need not concern us in the context of justification. Here, decision-analysis is used in order to ascertain whether it is possible to give a rational reconstruction of belief. For this purpose, we can treat cognitive actions as decisions. There are many cognitive actions, but for the present discussion, the cognitive actions will be restricted to those related to a set of theories $\mathcal{T}$ proposed to explain, describe, or predict some aspect of reality. The relevant decisions are then of the kind 'accept theory T over all other theories in $\mathcal{T}$'.

The next question is: what are the relevant goals? Again, we should be considering cognitive goals. Important epistemic aims are for instance truth, coherence, relevance, universality, and simplicity. It is questionable whether these epistemic aims exhaust the goals that drive actual scientific behaviour. More mundane goals such as achieving professional recognition, publishing articles in high-impact journals and rising in the academic ranks, undoubtedly affect the behaviour of scientists. Some sociologists of science argue that it is these latter-type goals that are the main impetus for any scientific activity (see Gustin, 1973, for an exposition and critique), and that epistemic

aims only enter as post-hoc justifications. Such radical views have been criticised for undermining the rationality of science. But in the present framework this is not necessarily the case. A scientist, motivated by a need for professional recognition, acting solely to promote this goal by stealing ideas, forging data, badmouthing the competition, may be considered a rational one. For such measures can, and maybe often do, result in awe and recognition if the particulars of the chosen strategy remain undetected. Yet, the ascription of the predicate 'rational' to such a scientist does not come readily. But I do not think this should be attributed to a wrong definition of 'rationality'. Rather, I find it hard to believe that someone can be solely driven by a goal of professional recognition. But if such a person exists, so be it. In the present framework, such a person should be called rational. However, such a person will not be called a rational *scientist*, if one takes the aim of knowledge to be a minimal defining characteristic of a scientist.

The final and probably hardest question is: what are the relevant consequences? Each action should have some value in realising the aspired aims. If one of the theories in $\mathcal{T}$ is taken to be true, only one action has the desired consequence of realising the goal of truth. If the theories in $\mathcal{T}$ differ in simplicity, it might be possible to order them along this dimension, so that one action (accepting the simplest theory) has the consequence of maximising the attainment of this goal, while the others serve this goal to a lesser extent (although all serve the goal to some extent). However, as will become clear in Chapter 3, simplicity is not a well-defined characteristic of a theory. This problem complicates ordering theories on the basis of simplicity. But, a subjective evaluation of simplicity is enough in order to render the choice between theories individually rational, as long as such evaluations can be made in a consistent manner[1]. Assigning epistemic consequences to theories, models, hypotheses, and such, is certainly not without problems. But that aims form the basis of an ordering of the available actions is essential to a decision-theoretic analysis. It is customary to call this ordering a utility ordering. The optimal decision is the action with the highest utility, or, if there is some uncertainty, the highest expected utility.

Uncertainty is always present in scientific inference. Otherwise, science would be an easy and rather dull enterprise. One way to view decisions under uncertainty is as a game. We can take scientists to play a game against nature. Nature has made a certain move (her state) and the scientist must make a counter-move, such as adopting a certain belief regarding nature's state. The utility (or payoff) of this move depends on nature's unknown state. So instead if assigning utilities to each move or action of the scientist, utilities are assigned to each pair of moves (one for nature and the other for the scientist). We'll call the situation just described an epistemic game $\Gamma \equiv \langle \Theta, \mathcal{A}, u \rangle$. The game consists of a set of moves for nature Θ, a set of actions for the scientist $\mathcal{A}$, and an utility function $u(\cdot)$ over the cartesian product $\Theta \times \mathcal{A}$. No utility function for nature is included, since I find it implausible that nature is a strategic player in the standard sense. To illustrate, let's consider a simple example in which nature has two possible states, θ_1 and θ_2, and the scientist two actions, a_1 (believe the state of nature to be θ_1) and a_2 (believe the state of nature to be θ_2). If

[1]One demand of consistency is for instance that the relative simplicity of T_1 to T_2 should not change when T_3 is added to (or removed from) the choice set.

		Nature	
		θ_1	θ_2
Scientist	a_1	α	$-\alpha$
	a_2	$-\alpha$	α

Figure 2.1: A simple epistemic game

we take the scientist to be concerned only with being right, then the utility matrix of the game might look something as in Figure 2.1. Without any further information, this game does not have an immediate solution. None of the actions has an overall higher utility. In other words, the choice for a particular action is underdetermined.

2.2 Underdetermination

Underdetermination is a well-known problem in the philosophy of science. The thesis of underdetermination can be traced back to Hume's (1748/1910) classic problem of induction. This problem concerns the confirmation of universal laws of the type 'all A's are B's'. As long as one cannot observe, or has not observed, all members of class A, there is no way in which this law can be conclusively confirmed. No matter how many black ravens one has observed, there is always a possibility that an unobserved raven does not have the property 'black'. No finite body of data can prove the correctness of a universal law, simply because a finite body of data cannot cover the infinite domain to which the law applies. So universal laws are underdetermined by empirical data.

Hume's thesis is a reminder that induction does not result in absolute certainty. Factual inference does not have the same status as logical inference. One may be able to arrive at theories which are probably true, but one cannot arrive at a theory which is necessarily true. A different slant on the underdetermination problem is given by Goodman (1954). His 'new riddle of induction' is introduced as follows. Consider the theory

T_1: All emeralds are green.

Confirmation of this theory comes from the observation of a number of emeralds, which were indeed all green. Hume's problem was that, unless one has observed all emeralds that have and will ever exist in the world, it cannot be ruled out that an emerald ever exists which is not green. Goodman's problem is of a different nature. He introduces another theory

T_2: All emeralds are grue,

in which grue is defined as

Grue: An object is grue if it is green before time t and blue after.

The problem is that, until time t, both theories have exactly the same prediction, namely that all emeralds will be observed to be green. So both theories are equally

supported by observations made before t. The theories differ only in their predictions of emeralds after t. After t, T_1 would still be confirmed by observing a green emerald, while this would disconfirm T_2. But, when t lies somewhere in the future, why should we prefer T_1 over T_2, if both are equally supported by empirical evidence?

One might argue that T_2 is a strange theory, since it introduces a colour definition which does not stick with our intuitive colour definitions. But that is just the point Goodman is trying to make. For us, being speakers of the green-language, it is strange to speak of 'grue'. But apart from habit, there is no reason to prefer terms such as 'green' to terms such as 'grue'. Another argument against T_2 might be that T_1 is a simpler theory, since it does not involve an extra time-dependence. To this Goodman would reply that, although T_1 is indeed simpler for someone who normally uses the terms 'green' and 'blue', it isn't for someone who uses terms like 'grue'. To see this, define 'bleen' similar to 'grue' as

Bleen: An object is bleen if it is blue before time t and green after.

Now, T_1 isn't immediately sensible to a member of the 'grue' language community. It needs to be translated as

T_1': all emeralds are grue before t and bleen after.

Comparing this formulation to T_2 shows that, for a member of the grue community, T_2 is simpler. Admittedly, Goodman's example is somewhat contrived. But similar problems do appear in actual science. Take Freud's notion of repression. I won't do justice to the full theory, and formulate it simply as:

Repressed: M is repressed if it could be remembered before t_1, could not be remembered after t_1 and before t_2, and could be remembered after t_2.

where t_1 denotes the time of repression and t_2 the time of some unblocking action by the therapist. Now define the theory of repression as

T_1: All ego-threatening states M are repressed,

and take the alternative theory

T_2: All ego-threatening states M are forgotten,

in which forgotten is defined as

Forgotten: M is forgotten if it could be remembered before t_1, but could not be remembered after.

The resemblance to the grue problem should be clear.

Goodman's new riddle of induction strengthens the case for underdetermination. For now there are multiple theories which are equivalent with empirical data. One way to deal with Goodman's and Hume's problem is to reject induction as a form of scientific inference. And if rejecting induction, why not reject confirmation at the same time? This was Popper's (1959) strategy. For while universal laws cannot be conclusively confirmed, they can be conclusively falsified, even by a single observation.

Or at least, that was Popper's idea. For there is a version of the underdetermination thesis which denies conclusive falsification as well as confirmation. This general thesis of underdetermination is related to the Duhem-Quine thesis. Duhem argued that, in order to empirically test a theory, one derives predictions from it which are compared to observable evidence. But the predictions do not follow directly from the theory itself. To derive predictions, the theory must be supplemented with auxiliary hypotheses. The problem then is that when the observed evidence does not agree with the predictions, this can mean that the theory is wrong, but also that the auxiliary hypotheses are wrong. This is an instance of the modus tollens rule in predicate logic. If a theory T and auxiliary hypotheses A together imply an observable event E, i.e. $(H \wedge A) \implies E$, then the valid conclusion from not observing the event is $\neg E \implies \neg(T \wedge A)$. But $\neg(T \wedge A)$ means that either $\neg T$ or $\neg A$ (or both). So when predictions are disconfirmed, this is not necessarily a disconfirmation of the theory, but only of the combination of the theory and the auxiliary hypotheses. Empirical tests can only test the whole of theory and auxiliary hypotheses. As such, there are no critical tests of theories. Theories are neither confirmed nor falsified by empirical evidence. Quine calls this position 'holism'. As a reaction to the disconfirmation of predictions, one can choose to abandon the theory or the auxiliary hypotheses. The data do not give precedence to one of these actions. The crux of the Duhem-Quine thesis is that it is always possible to adjust the auxiliary hypotheses in such a way that the combination of theory and adjusted auxiliary hypotheses is consistent with the observations. If this is the case, then any theory can be saved in the face of disconfirmation. This led Quine (1951, p.40) to the often cited maxim that: 'Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system.'

As Quine (1975) points out, the thesis of holism is different from the strong thesis of underdetermination which is sometimes attributed to him. Holism refers to a type of underdetermination, in the sense that the choice for which aspect in a network to adjust, in order to reconcile disconfirming evidence, is underdetermined. The strong thesis of underdetermination refers to underdetermination of whole networks. It takes that for any given network or 'total system of the world', there exists a rival which is empirically equivalent. The empirical equivalence is not bounded to a given body of data, but extents to any possible body of data, to all possible observations. Moreover, while empirically equivalent, the theories are logically incompatible and cannot be made logically equivalent by a reconstrual of predicates. This is a very strong thesis indeed, and Quine remains somewhat vague about his commitment to it. But it is this thesis which leads to the rejection of realism as a viable position (e.g. Van Fraassen, 1980).

2.3 Back to decisions

The simple epistemic game described in Section 2.1 was of course overly simple. Its purpose was mainly to provide a smooth transition to the topic of underdetermination. Since it has served that purpose, I shall now extend it into a more realistic version. One important omission was the scientist's belief regarding the probability that nature

		Nature	
		θ_1	θ_2
Scientist	a_1	α	$-\alpha$
	a_2	-3α	$\frac{1}{3}\alpha$

Figure 2.2: An epistemic game with simplicity

is in a particular state. Such belief would allow the scientist to choose the action with the highest *expected* utility. If the scientist can assign a non-identical probability to each state, the underdetermination may vanish. Say that state θ_1 is assigned a probability of .25, and state θ_2 a probability of .75. Then $E[u(a_1)] = .25\alpha - .75\alpha = -.5\alpha$ and $E[u(a_2)] = .75\alpha - .25\alpha = .5\alpha$, so that a_2 is the rational strategy. If the probability assignment is entirely subjective, made without regard to anything independent of the scientist, we have a case of pure armchair research. While there is nothing inherently wrong with maximising subjective expected utility, it is more common that the scientist has empirical evidence which is informative (although not decisive) of the state of nature. Ideally, this information would be fully captured by a probability distribution over Θ. In that case, the scientist can proceed as before, by computing the expected utility of his actions and adopting that action with the highest expected utility. Unfortunately, the informational value of empirical evidence for the state of nature is not a simple matter. Some problems will be addressed in Chapter 3. From the strong underdetermination thesis, it follows that empirical evidence can have an identical informational value for indefinitely many actions. Regarding underdetermination, two reservations are in place. Firstly, the thesis concerns sets of 'proposable' theories, not sets of actually proposed theories. The requirement that a scientist can justify his decision against all other possible decisions may be overly strong. If the requirement is that the decision can be justified against a set of decisions under actual consideration, then the problem of underdetermination may not be of much practical importance. For while there exists an empirically equivalent theory for every theory under consideration, this equivalent theory may not itself be one under consideration. Secondly, the thesis in its classical formulation concerns empirically equivalent actions. Some authors (e.g. Laudan, 1996) have argued that the problem can be avoided when other goals than empirical adequacy are considered. But, although initially plausible, it is not that easy to sweep underdetermination under the carpet. To see this, consider a situation in which the scientist has additional aims to being right. For instance, the scientist values simple theories over more complex ones. If θ_2 is more complex than θ_1, the utility matrix might look something like Figure 2.2. Again, the game does not have an immediate solution, since no action dominates the other (an action is dominant if its utility is at least as high as the utility of the other actions for every element in Θ). Now consider the case where the scientist has some information regarding the probability of the different states of nature. The game is supplemented by a probability function over Θ, and if this function is the same as before, then $E[u(a_1)] = -.5\alpha$ and $E[u(a_2)] = -.5\alpha$. So in this case the scientist should still be indifferent to the two actions. Hence, the choice

for an action is underdetermined.

When there are multiple goals to realise, we enter the field of multi-attribute decision making. A simple form of multi-attribute utility is additive. If the consequences of an action can be described by a vector of values of the particular action (relative to the goals), then under additivity, the utility of the action can be stated as $u(a) = \sum_j w_{ij} v_j$, in which w_{ij} denotes the relative weight given to attribute j by individual i. It is easy to see that two actions, a_1 and a_2, with different values, $v_1 = (v_{11}, v_{12})$ and $v_2 = (v_{21}, v_{22})$, may have the same utility. Suppose that an individual weights the importance of achieving the goals as $w_i = (.75, .25)$, then two actions with respective values of $v_1 = (4, 1)$ and $v_2 = (2, 7)$ have an equivalent utility. So undecidability remains a potential problem even if more aims are added. I will call theories with identical epistemological utilities axiologically equivalent:

Definition 1 (Axiological equivalence).
Theories $T \in \mathcal{T}$ are axiologically equivalent relative to a set of goals $\mathcal{G}$ if $(\forall T_i, T_j \in \mathcal{T})[u_{|\mathcal{G}}(T_i) = u_{|\mathcal{G}}(T_j)]$.

Axiological equivalence leads to the notion of axiological underdetermination:

Definition 2 (Axiological underdetermination).
Given a set of goals $\mathcal{G}$, the choice for a theory T is axiologically underdetermined if $(\forall T')[u_{|\mathcal{G}}(T') \preceq u_{|\mathcal{G}}(T)]$ *and* $(\exists T' \neq T)[u_{|\mathcal{G}}(T') = u_{|\mathcal{G}}(T)]$.

Underdetermination in the classical sense is subsumed in axiological underdetermination. If empirical adequacy is the only aspired aim, the two types of underdetermination are identical. Note that an axiologically underdetermined theory is not necessarily empirically underdetermined. Other aspects of theories can be traded off against empirical fit in such a way that the epistemic utility of two theories is identical. In the previous example, this was shown for the trade-off between simplicity and empirical fit. So contrary to earlier claims, the addition of aspired aims can actually raise the probability of underdetermination, for there are more ways in which trade-offs can be realised.

2.4 Relativism, conventionalism, and social constructivism

Responses to challenges such as underdetermination take a variety of forms. One, which will not be discussed extensively, is to deny that there is a real problem. Problems such as underdetermination simply pose limits on what science can reasonably achieve. Van Fraassen's (1980) position, which he calls constructive empiricism, is that science should stay within these limits and focus solely on empirical adequacy. If there is a pluralism of perspectives which achieves this, so be it. There is no need to designate one as the best. Constructive empiricism is a normative position, stating that empirical adequacy is all that scientists should be concerned with. In practice however, scientists do want to go further. Moreover, they wish to justify such moves as non-arbitrary. So there is a mismatch between the goal prescribed by constructive empiricism and the goals actually held by scientists. Many scientists are of a realist persuasion, whether outright or secretly. Realism is the position that theoretical

terms refer to things in reality. Theories can be true or false, and the truth of a theory depends on events in reality. The realist aim is to arrive at such true theories. Most will realise that it is a difficult goal to achieve, and that it is even more difficult to recognise when the goal is actually achieved and when not. But this does not render the goal of realism irrational. Goals themselves are not subject to designations of rationality, only acts of striving to realise a goal are rational or irrational. When goals are conflicting, so that they cannot all be obtained concurrently, then realising one is irrational if unrealised other goals are deemed more important. Now, I don't believe that the goal of realism is conflicting in such a way with other epistemic aims such as empirical adequacy. But even if it was, waving aside the realist *aim* as irrational would still be nonsense.

Instead, one might argue, a la Van Fraassen, that the realist aim is inconsequential, since there is no way in which it can be overtly obtained. That would require means of conclusively separating true from false theories, and those means are not available. Although scientists may be driven by a realist aim, they cannot invoke this aim to justify their decisions[2]. The three -ism's in the title of this section, conventionalism, relativism and social constructivism, take such a position. Contrary to constructive empiricism, they do not take underdetermined choices as entirely arbitrary. While particular choices are arbitrary from a purely logical viewpoint, they are not arbitrary to an individual scientist or scientific community. The three -ism's are varieties of subjectivism (indeed another -ism), in the sense that their central concepts are not devoid of reference to a particular person or group of persons. Their tenets are quite similar, and it is difficult to make a clear distinction apart from their genesis.

Conventionalism was touched upon at the start of this chapter as a possible strategy to avoid infinite regress in justification. Conventionalism takes truth (or at least some truths) to be a matter of convention. The logical positivists took this position to solve the problem that logical truths depend on the truth of logic, while the truth of logic is neither a logical nor empirical truth. So the truth of logic is established by convention. A different conception of conventionalism stems from the work on the foundations of geometry, most noticeably by Poincaré (1905/1979). Realising that the choice for a specific geometry, such as Euclidian geometry, is underdetermined by empirical data, Poincaré argued that this choice constitutes a convention. This is not to say that it is an arbitrary choice. According to Reichenbach, conventions such as a particular geometry are 'co-ordinate systems'; they are general rules which are laid down so that the terms and concepts in a theory have a well-defined subject-matter (Hibberd, 2001). While they have an a priori status, this is not in the sense of Kant's

[2]To scientist A's claim that 'I believe X because I strive for truth, and X is true', a sceptic can always respond 'How is X true?'. In other words, A must justify this claim. He might do so by stating that X has ample empirical support, but then he should also offer means to separate true from false theories on the basis of empirical support (i.e. some point p below which theories are false and above which theories are true). And even if A was able to do so, his actions would be identical to those of scientist B who strives for theories with empirical support above p. For justification, as long as 'truth' is not a directly inferable property, the realist aim can always be replaced by different aims, reflecting the indicators of truth. Hence, in the context of justification, the realist aim can be deemed inconsequential. If empirical support is directly inferable, scientist B may find himself in a more comfortable position, since his goal can then be overtly obtained. Insofar as truth is not identical to 'empirical support above p', scientist A will still be uncertain whether he actually achieved his goal. But, on another note, who expects certainty?

synthetic a priori knowledge. They are not for all time and independent of experience, but rather 'before knowledge'. Without them, there simply is no knowledge possible.

Relativism can be taken as merely another term for subjectivism. Indeed, relativism is such a broad and loosely jointed myriad of positions, that this would be correct. The relevant subspecies of relativism here is epistemological relativism, the position that the status of a knowledge claim can only be determined relative to the individual or group making that claim. Feyerabend (1975) is a well-known epistemological relativist. Taking knowledge in the classical sense as justified true belief, one can take either justification, truth, or both to be relative. Truth-relativism is a difficult position to defend, for it is self-refuting. If all propositions have different truth-values, depending on one's circumstances, then so does this position. A weaker version of truth-relativism is that some propositions may have a truth-value in one framework, but not in another, because the proposition has no meaning in that framework. This is a short characterisation of the thesis of incommensurability, as forwarded by Kuhn (1970) and Feyerabend (1975). Incommensurability renders a rational choice between competing frameworks impossible, for there is no common standard by which to compare different frameworks. Kuhn (1970) calls general scientific frameworks paradigms, and argues that the move from one paradigm to another has to be based on a 'leap of faith', rather than a rational decision. A little later, Kuhn (1977) admitted that there are paradigm-overarching principles that can be used to guide paradigm choice. But such methodological principles are too vague to determine paradigm choice. Feyerabend took a stronger position. He argued that there are no methodological rules which have not been profitably broken in the history of science. So there are no methodological rules that, upon their adoption, render a particular inference superior to another. The only defendable methodological principle is that 'Anything goes'.

While the conditionals in conventionalism and relativism can be either individuals or groups, social constructivism is explicitly a collectivist thesis. In its most radical form, it claims that all facts are constructions of 'thought collectives'. Facts are not discovered, but decreed through a social process of argumentation and negotiation. As we have seen, there are definite problems in justifying a theory choice by stating that it meets the realist aim. From this problem, social constructivists seem to infer that it is an impossible aim to be met (and not just impossible to show that it has been met). Scientific theories do not reflect reality in any way, but rather construct a 'reality'. That there really is no objective reality out there is a radical, and I think radically foolish, thesis. In order to understand this claim, one should realise that social constructivists equate 'truth' with 'generally accepted as true'. That is, they conflate 'X is true' with 'X is believed to be true' and as such hold a consensus theory of truth (Fine, 1996). According to social constructivists, 'truth' and 'fact' are not properties of objects or phenomena 'out there', but properties of social groups. Social discourse determines what is true or false, or fact or fiction. Reality plays no role in such matters.

I shall defer further discussion of the three -ism's to Chapter 6. The short characterisation offered here should suffice to realise that they offer a similar solution to the underdetermination problem. Namely, that while theory choice is underdetermined by empirical evidence, it is determined by empirical evidence *and* additional factors.

The additional factors being conventions (conventionalism), scientific frameworks or paradigms (relativism), or social processes such as argumentation and negotiation (social constructivism).

2.5 No place for normative methodology?

Relativistic arguments are often employed to reject the normative status of methodological rules (e.g. Feyerabend, 1975). Normative rules are imperatives, and with Kant we should distinguish between categorical and hypothetical imperatives. Categorical imperatives apply without qualification. For instance, 'You should not kill' is a categorical imperative. Hypothetical imperatives, on the other hand, are of the form 'If you want to achieve G, then do A'. For instance, 'If you want to live life outside of prison, you should not kill' is a hypothetical imperative. Methodological rules are hypothetical imperatives. Since methods are instrumental, they serve a certain goal. Clearly, the imperative applies to people striving for that goal. There is no reason to deny the normative status of methodological rules as hypothetical imperatives. Of course, it is up to the methodologist to show that (and where possible how) methodological rules forward certain aims. This is difficult, to be sure, but I don't think it should be considered impossible in principle.

The position that follows from the decision-theoretic framework is relativistic in a certain sense. The justification of scientific inference is relative to aspired aims. I see no problem in such a form of relativism. Under certain circumstances, it is equivalent to a relativism in Kuhn's sense. If different paradigms entail completely different sets of aims, then justification is relative to a paradigm. But many scientific aims, such as empirical adequacy and coherence, are paradigm-overarching. Some aims may even be pursued by all scientists. In the current framework, foundationalism in epistemology can be seen as the search for such universal aims. For if aims are universally shared, justification relative to these aims is universally valid.

2.6 Consensus

As soon as the conditional in relativism is not a single individual, but a group, consensus enters the picture. For instance, the justification of belief as knowledge may be taken as dependent on consensus. It has also been argued that not only the justification, but also the formation of belief, depends on consensus. For one thing, most of our knowledge does not stem from direct experience, but rather others' experiences that have been communicated to us. An important factor in this social process is the trust that is placed in others' knowledge claims. A way to determine the validity of such claims is by determining whether there is consensus between others on the claim. While the idea that much of our knowledge was transmitted by others is still compatible with the view that knowledge ultimately results from direct experience (albeit now someone else's experience), it has been argued that experience itself is determined by social processes. For instance, Vygotsky (1978) argues that higher psychological processes, such as thinking, arise from the internalisation of social interaction. In a similar vain, Levine, Resnick, and Higgins (1993) state that 'all mental

activity – from perceptual recognition to memory to problem solving – involves either representations of other people or the use of artifacts and cultural forms that have a social history' (p.604). A common argument for the social determination of knowledge is that language is a social medium, and experience and thought draw upon language to make them interpretable. Hence, experience and thought are mediated by social factors. Such ideas can be classified as 'social constructivist', although they differ from radical social constructivist accounts of science. But consensus is not only an important concept for scientists of a relativist persuasion. The history of science is often portrayed as a movement from consensus to consensus (e.g. Kitcher, 1993; Kuhn, 1970). In this sense, the key advancements in a scientific field are identified by consensus. On another note, some form of consensus may be deemed necessary for science to function at all. For instance, Popper's (1959) falsificationism can only work if there is consensus on what the basic statements are. For a test of a hypothesis is a critical test only to the extent that evidence is universally accepted as critical evidence for or against the hypothesis. In this sense, consensus can resolve the Duhem-Quine problem. For if there is universal consensus that the auxiliary hypotheses are true, then a critical test of an isolated theory is possible.

Ziman (1968, 1978) assigns a higher status to consensus, taking it to be the first principle of science. Attempting to demarcate science from non-science, he argues that the only principle on which such a distinction can be made is that science strives for a rational consensus. In the current framework, the incorporation of consensus in the set of aims would mean that the methods of science should realise this aim. According to Ziman, the methods commonly used by scientists serve just this aim. De Groot (1961, 1982) takes this idea a step further. Instead of justifying scientific methods (partly) by arguing that they implicitly serve the goal of consensus, he proposes a normative methodology which explicitly promotes the goal of consensus. This idea, which is the key to his Forum-theory, is the topic of Chapter 6. The intermittent chapters concern the relation between consensus and the issue of underdetermination, as well as social information integration and collective decision making. But before moving to these topics, the meaning of the term 'consensus' should be addressed.

2.6.1 A definition of consensus

Often, the term 'consensus' is employed as a proper synonym for agreement between people (Horowitz, 1962). So, there is consensus in a group that X if all individuals in the group agree that X. But, while surely a necessary aspect of consensus, this is not the whole story. To see why there must be more to consensus, consider a case of 'false dissensus'. There is false dissensus if everyone believes X, but nobody believes anybody else to believe X. For clarity of presentation, especially later on, the belief-operator B_i will be introduced, which is taken to mean 'individual i believes ...'. In the situation just described, there is a group of individuals $\mathcal{P} = \{1, \ldots, n\}$, who all believe X

$$(\forall i \in \mathcal{P})[B_i(X)],$$

while each individual believes the other's don't believe X

$$(\forall i \in \mathcal{P})(\forall j \neq i \in \mathcal{P})[B_i B_j(\neg X)].$$

Does it make sense to speak of 'consensus' in such a situation? I think not. This situation of false dissensus is one of 'accidental' agreement, but not consensus. Consensus at least requires that the agreement is recognised. So we may speak of consensus if

$$(\forall i, j \in \mathcal{P})[B_i B_j(X)],$$

that is, if each individual believes all individuals (including him- or herself) believe X. I shall take it for granted that if a person believes he/she believes X, he or she actually believes X (i.e. $B_i B_i(X) \implies B_i(X)$). The demand for recognition should be extended further, since it is strange to speak of consensus if the recognition of agreement is itself not recognised. That is, if everyone believes X, everyone believes everyone else believes X, but no one believes everyone else believes everyone else believes X:

$$(\forall i \neq j, k \in \mathcal{P})(\forall j, k \in \mathcal{P})[B_j B_k(X) \wedge B_i \neg B_j B_k(X)].$$

Since the requirement of recognition should be reapplied to each recognition, that is, to all higher-order beliefs, we arrive at the notion of common belief, akin to that of common knowledge (Lewis, 1969). Common knowledge or belief is an important concept in game theory (Aumann, 1976; Geanakoplos, 1992), the field of multi-agent systems (Fagin, Halpern, Moses, & Vardi, 1995), and theories of meaning (Schiffer, 1972), and discourse understanding (Clark & Marshall, 1981). There is common belief in a group if everyone believes X, believes everyone believes X, believes everyone believes everyone believes X, etc. For a concise definition, define the operator $E_{\mathcal{P}}$ as 'everyone in $\mathcal{P}$ believes ...', i.e.

$$E_{\mathcal{P}} X \equiv \bigwedge_{i \in \mathcal{P}} B_i X.$$

Then $E^k_{\mathcal{P}}$ can be defined recursively as

$$\begin{aligned} E^1_{\mathcal{P}} X &\equiv E_{\mathcal{P}} X \\ E^k_{\mathcal{P}} X &\equiv E_{\mathcal{P}} E^{k-1}_{\mathcal{P}} X \quad \text{for } k > 1, \end{aligned}$$

and consensus as common belief as:

Definition 3 (Consensus).
There is consensus in a group $\mathcal{P}$ that X if $E^{\infty}_{\mathcal{P}} X$.

2.6.2 Attaining consensus

The formal definition of consensus given above requires an infinite series of beliefs. But such an infinite series of beliefs cannot be actually entertained in practice (if believing something takes time, no matter how little, someone can only entertain a finite number of beliefs in a finite lifetime). Moreover, there is no way to provide

conclusive empirical evidence for the existence of this consensus, for this would require verifying the existence of an infinite number of beliefs. Scheff (1967) and Bach (1975), who arrived at a similar definition of consensus, argue that there is no need for an infinite series, and that $E^3_{\mathcal{P}}$ is enough to speak of consensus. However, the level of agreement necessary differs from case to case. Truncating the series is certainly not always a viable solution. In many situations, such as the problem of 'coordinated attack', supposed disagreement on a higher level will lead people to behave as if there was no agreement at all (for another example, see Rubinstein, 1989). A better solution, offered by Lewis (1969), is that the infinite series should be viewed as a chain of implications, not as actual steps in someone's reasoning. Such an infinite chain of implications can be based on a few basic premisses. According to Lewis, there is common belief in a group $\mathcal{P}$ that X if a state-of-affairs S holds, such that

(a) Everyone in $\mathcal{P}$ has reason to believe that S holds.

(b) S indicates to everyone in $\mathcal{P}$ that they all have reason to believe that S holds.

(c) S indicates to everyone in $\mathcal{P}$ that X.

What kind of state-of-affairs is S, and how does the chain of implications proceed? I will try to make the idea more tangible by an example. It seems a safe assumption that, amongst all people with at least a primary education, there is consensus that the earth is round. Why? Because being taught that the earth is round is part of all primary education. So every individual with a primary education believes that the earth is round. Denoting the set of people with at least a primary education as $\mathcal{P}$, and the proposition 'the earth is round' as X, then

$$E^1_{\mathcal{P}}\colon\ (\forall i)[(i \in \mathcal{P}) \implies B_i X].$$

Whether I believe that everyone with a primary education believes the earth is round rests on my belief in E^1. But in order to believe that everyone believes everyone else believes the earth is round, I should believe that everyone else in $\mathcal{P}$ also believes E^1. A possible justification for this belief is that membership of $\mathcal{P}$ implies belief in E^1

$$E^2_{\mathcal{P}}X\colon\ (\forall i)[(i \in \mathcal{P}) \implies B_i E_{\mathcal{P}} X].$$

In order to believe the third-order shared belief $(\forall i, j, k \in \mathcal{P}) B_i B_j B_k E$ a similar assumption can be made, i.e.

$$E^3_{\mathcal{P}}X\colon\ (\forall i)[(i \in \mathcal{P}) \implies B_i E^2_{\mathcal{P}} X],$$

and so forth. So a chain of reasoning is instigated from 'being a member of $\mathcal{P}$'. Being a member of $\mathcal{P}$ implies that one has certain beliefs, as well as certain beliefs about the beliefs of other members of $\mathcal{P}$. So essentially

$$\mathcal{P} \implies E^\infty_{\mathcal{P}} X.$$

But how does membership of $\mathcal{P}$ imply such consensus? A possible answer is that 'believing X' is a defining characteristic of the social-identity of group $\mathcal{P}$. If this social identity is understood by all members of $\mathcal{P}$, and all members of $\mathcal{P}$ share an

implication-schema, so that $(\forall i \in \mathcal{P})(E_{\mathcal{P}}X \implies B_i E_{\mathcal{P}}X)$, then consensus itself is implied.

Another and easier way in which consensus can be reached is by overt communication. If all members in a group come together and voice their belief that X, and nobody has reason to suspect someone to falsely report this belief, then there is reason enough for everyone to believe everyone believes X. Moreover, there is reason enough for everyone to believe there is reason enough to believe everyone believes X, etc.

While infinite series of beliefs are counter-factual and impossible to confirm empirically, the definition of consensus given here is not entirely useless. For consensus is possible when taken as a chain of implications, based on for instance group-identity or overt communication. Also, the practical relevance of higher-order beliefs often diminishes as we move higher up in the hierarchy. Let's go back to the Duhem-Quine problem of critical tests. Second-order shared belief in $\mathcal{P}$ of the auxiliary hypotheses A (e.g. $E^2_{\mathcal{P}}A$) seems enough in order for everyone in $\mathcal{P}$ to accept a test of theory T as a *critical* test for all members of $\mathcal{P}$. However, for everyone to accept that the test is a critical test for everyone in $\mathcal{P}$ would require $E^3_{\mathcal{P}}A$, a third-order shared belief.

2.6.3 Implications of consensus

Consensus (or common belief) is often taken as necessary prerequisite for interpersonal coordination (Lewis, 1969; Bach, 1975). An example of this necessity is given in the problem of 'coordinated attack' (Fagin et al., 1995). Suppose there are two generals, who are stationed with their armies on hills located on opposite sides of a valley, with the enemy army between them. The generals must attack, but an attack will only be successful if both generals attack at the same time. If any general would attack by himself, he would suffer unacceptable loss. The only means of communicating with each-other is through messengers who might not arrive at their destination, since they have to travel through enemy territory. So the generals cannot be sure that their message was actually delivered to the other. Suppose that general A plans on attacking at dawn, and sends a message to general B to inform him about his intentions. Since general B knows that general A will only attack if he is sure that general B will do so also, he sends a messenger back to acknowledge his agreement with the plan. However, general B cannot be sure that this message arrived. Since general B will only attack if he is sure that A does so also, and A knows this, A sends a messenger back acknowledging the receipt of B's message. But for A to attack, he must be sure that this message has arrived, which would require another message back, and so forth. There is no way in which both parties can coordinate the attack.

The coordinated attack problem is a clear example in which common belief is necessary. There is no point at which the infinite series of shared beliefs can be truncated while maintaining the effect of common belief. But clearly, the coordinated attack is rather different than situations encountered in scientific practice. Are there coordination-problems in science which might require consensus? This question will be addressed in Chapter 6. But to round off this chapter, here is one example where consensus might play a role in science: assumptions. It is not often that one finds an article in which all assumptions on which its results rest are stated explicitly. A possible reason for this is that doing so would diminish the persuasive force of the

argument in the article. For the number of necessary assumptions can be quite large, and many assumptions, if not inherently untestable, remain untested nevertheless. Explicit exposition of all untested assumptions renders the argument an easy target for criticism. Of course, not mentioning assumptions does not provide immunity to such criticism. A clever commentator could pinpoint the implicit assumptions taken. Consensus on the (plausibility of) assumptions however would 'immunise' the argument in a practical way, since no one would be able to criticise these assumptions without in some sense criticising him- or herself. So consensus on assumptions results in the absence of criticism to an otherwise criticisable assumption.

Consider as an example the assumption of a normally distributed variable. Since most hypothesis tests in psychology are either based on a t-test or F-test, this assumption is usually made. As once remarked by Lippman, the reason the assumption is so widespread is that 'Experimentalists think that [the Normal distribution] is a mathematical theorem while the mathematicians believe it to be an experimental fact.' (cited in Wright, 2003, p.128). Now this is an example of an assumption which is testable, for instance by a Kolmogorov-Smirnoff test. But nobody seems to take the effort to do this, or at least nobody mentions it when they do so. Of course, both the t-test and F-test are rather robust when it comes to violation of the normality-assumption. Also, because of the central limit theorem, there can be good reasons to assume normality. But still, testing the normality assumption takes very little effort nowadays. So why does nobody do it? There are two options that I would like to mention. The first is that the assumption is actually tested, but the test is not mentioned, because the assumption could not be rejected. If there is consensus that testable assumptions should always be tested, but the test should only be mentioned when it leads to a rejection of the assumption, this is fine. Any reader who is part of this consensus could then infer that the assumption holds. It is however questionable that this particular consensus exists. The second option is that there is consensus that the assumption is plausible and need not be tested. In this case, a researcher can quite safely proceed without testing, for neither himself nor peer reviewer or reader would find the results less convincing because of the absence of a test of the assumption. In this way, the normality assumption achieves the status of a priori knowledge. I think this is in general how consensus affects propositions: it obliviates the need to test them. Of course, this is just in the same practical sense that there is only a problem of justification if there is a disagreement in view.

3
The hierarchical theory of justification and statistical model selection

3.1 Introduction

The possibilities of language are endless. As Chomsky (1957) showed, natural language allows for the formation of an infinite number of grammatically correct sentences. For one thing, this is due to the fact that there is no theoretical limit to the number of dependent clauses that can be embedded in a given sentence, as in 'the rat ran', 'the rat the cat chased ran', 'the rat the cat the dog teased chased ran', etc. There is an infinitude of possible sentences which convey a certain message, and we can conceive an uttered statement as the result of an implicit choice from this infinitude. Now, consider a situation in which a scientist provides an explanation for a certain phenomenon. Just like the grammatically correct sentences, the number of possible explanations he or she can give is also infinite, since it will always be possible to add another explanans to an existing explanation. For the field of statistical modelling, the situation can be exemplified in what is known as the curve-fitting problem. In this regression-type problem, the goal is to find a curve that describes a relation between a set of paired observations (x, y), for instance by finding the function $f(x)$ that best predicts the values of variable y. Since, for any set of n paired observations there exists a $(n-1)$-degree polynomial that passes through all n points, an $(n-1)$-polynomial provides a perfect description of the occurrences of the data. Obviously, considering this maximal descriptive accuracy, an n-degree polynomial passing through all points is an equally good candidate for the unknown function, and so is an $(n+1)$-degree polynomial, an $(n+2)$-degree polynomial, etc. For any finite set of paired observations (x, y), there is a family of best fitting curves which has infinite members. When based purely on the accuracy of description, the choice for one from

the family is arbitrary.

Science does not pride itself on the arbitrariness of its explanations; the choice for a certain explanation must be motivated and this motivation should be explicit and scientifically valid. However, this problem of justification has not been adequately solved for the empirical sciences. Without any objective criteria for justification, it is quite reasonable for scientists to disagree, or at least, not irrational to do so. Still, scientists do reach agreement, and quite often so. Since this consensus is not the logical result of a reference to the universal rules of scientific inference, there must be another explanation. One of these involves the postulation of what can be called the hierarchical theory of justification. It is also known as the theory of instrumental rationality, and its influential advocates include Popper (1959), Hempel (1965), and Reichenbach (1938), amongst others.

3.1.1 The hierarchical theory of justification

According to the hierarchical theory of justification, there are three interrelated levels at which, and by means of which, consensus is forged (Laudan, 1984). The factual level is the lowest in the hierarchy and concerns matters-of-fact. In this context, the term 'matters-of-fact' refers to all descriptions and explanations of what there is in the world, formulated as hypotheses, models, theories, etc. The hierarchical theory prescribes that disagreement concerning such matters-of-fact can be resolved on the methodological level, which is one step up in the hierarchy. This level consists of the rules concerning empirical support and theory comparison. The methodological rules constituting this level will not be the universal laws of scientific inference as sought by the logical positivists, but rather the inter-subjective rules that are part of specific paradigms (e.g. Kuhn, 1970). It may be possible that scientists do not agree on the proper methodological rules. The hierarchical theory prescribes that disagreement on the methodological level must be resolved by moving to the highest level in the hierarchy. This is the axiological level, which concerns the goals and aims of science. These values may resemble the Mertonian norms of science (commonality, universalism, disinterestedness and organised scepticism) but other, more specific values can be considered. In fact, any values by which to judge the merits of theories might be considered here. The model allows for disagreement on an axiological level, but this disagreement cannot be resolved on a higher level in the hierarchy and so will remain unresolved. A summary of the hierarchical theory in graphical form is given in Figure 3.1.

As a preliminary example, let's impose the hierarchical theory onto the curve-fitting problem. The problem is to provide an explanation for occurrences in a given dataset by specifying the generating function that could have resulted in the observed data. Consider two scientists who both have found 'the' explanation for the data: scientist A proposes an $(n-1)$-degree polynomial and scientist B an n-degree polynomial; how can the scientists resolve their disagreement? A reasonable solution would be to compare their explanations on a common criterion and decide which is optimal. That is, they must agree on a proper method to evaluate the explanations and decide on the basis of the result of this evaluation. Clearly, the choice for such an evaluation method should depend on what both scientists require from an expla-

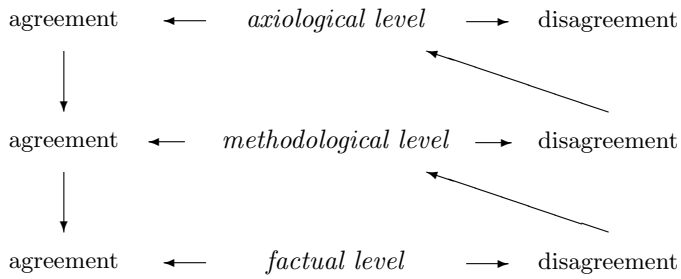


Figure 3.1: The hierarchical theory of justification

nation. One requirement already mentioned may be the precision of the explanation, which in this case may be defined as the fit between the proposed model and the observed data. Since both polynomials describe the observed data perfectly well, this criterion is not sufficient to distinguish the explanations. Another requirement may be that the explanation be as simple as possible. It will become clear later that this requirement may be problematic, but for now let's assume that the scientists agree that the number of degrees of the polynomial is a good measure of the simplicity of the explanation. Comparing both functions on this criterion, the scientists will agree on the optimality of the $(n-1)$-degree polynomial as the explanation for y. Recapitulating: the disagreement on the factual level concerning the best explanation for y was solved by evaluating the explanations with the agreed-upon parameter count method, whose use was motivated by the agreement on the axiological level that simplicity is a required value for any explanation.

3.1.2 Statistical model selection

The curve-fitting problem above is a simplified version of a situation that may arise when scientists disagree on the best statistical model. For statistical modelling, the problem of finding the best curve describing paired observations becomes the general problem of finding the best model to describe regularities in a body of data. A statistical model M is defined as a family of (multivariate) probability distributions $f(\boldsymbol{x}|\boldsymbol{\theta})$, characterised by parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ (Linhart & Zucchini, 1986). The observed data, consisting of n, q-variate observations $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, are considered to be the result of random sampling from a population which is governed by the data generating process $F(\boldsymbol{x})$. If $\mathcal{F}$ is the set of all q-variate distribution functions, a model $M(\boldsymbol{\theta})$ is a subset of $\mathcal{F}$, and each element in $M(\boldsymbol{\theta})$ is a fully specified model. A fitted model $M(\hat{\boldsymbol{\theta}}_x)$ is a fully specified model of which the parameters $\hat{\boldsymbol{\theta}}_x$ are fixed so as to minimise the discrepancy between the model and the observed distribution $f(\boldsymbol{x})$, as in maximum likelihood estimation. A best approximating model $M(\hat{\boldsymbol{\theta}}_F)$ is a fully specified model of which the parameters $\hat{\boldsymbol{\theta}}_F$

are fixed so as to minimise the discrepancy between the model and the data generating process. A model M is correctly specified if $F(\boldsymbol{x}) \in M$, and otherwise misspecified. A model M_1 is nested under model M_2 if $M_1 \subset M_2$, and these models are strictly non-nested if $M_1 \cap M_2 = \emptyset$.

The general problem of statistical model selection considered here is that of a set of models $\mathcal{M} = \{M_1, M_2, \ldots\}$ which are proposed to describe the data generating process $F(\boldsymbol{x})$. The objective is to choose the element $M^* \in \mathcal{M}$ that optimally represents $F(\boldsymbol{x})$. The elements of $\mathcal{M}$ may be nested or non-nested and correctly specified or misspecified. Moreover, the nontrivial situations are considered where $F(\boldsymbol{x})$ is not directly observable, so that the merit of each model as representing $F(\boldsymbol{x})$ can only be estimated. In practise, most of the parameters θ will represent substantial hypotheses concerning the data generating process, but some may represent the auxiliary assumptions necessary for a model's identifiability. Most researchers will agree that the selection of M^* should be more dependent on the agreement between substantial relations in the model and the population, than on the agreement between the ad hoc assumptions and the population (e.g. Golden, 2000).

3.1.3 Requirements of the hierarchical theory

According to the hierarchical theory of justification, disagreement over M^* can be resolved by reaching agreement on the proper methods to determine the models' merit, while disagreement regarding these specific methods can be resolved by reaching agreement on what qualities a method should address. The hierarchical theory of justification offers a normative theory of rational scientific decision-making and is an application of the decision-theoretic framework described in Chapter 2. For statistical model selection, the decision constitutes a choice from the model set $\mathcal{M}$ and this choice is to be made in accordance with epistemic aims or goals $g \in \mathcal{G}$. In other words, the choice for a model can be evaluated in terms of the consequences that choice has for the realisation of the epistemic aim g. If there are J aims under consideration, we may assume that each model can be characterised by J variables v_j, each reflecting the value of the model relative to the epistemic aim g_j. Not every aim may be deemed of equal importance to each decision-maker. As such, the values v_{jk} of a model k may be weighted by the relative importance w_{ij} of aim g_j to decision-maker i. If the values v_{jk} and the weights w_{ij} were directly accessible and both measured on an interval scale, the model selection problem would be solvable by standard techniques for multi-attribute decision making (e.g. Keeny & Raiffa, 1976). For instance, a simple and widely used decision rule is to choose the option which maximises the weighted sum, i.e. $M^* = \arg\max_k \sum_j w_{ij} v_{jk}$. Although some of the values v_{jk} may be directly inferred from the model M_k, often this will not be the case. When some of the values must be estimated on the basis of the model and the data, the decision procedure is more complicated, because a model which has the highest estimated worth may in reality have less worth than another model in the choice set. Any way, in statistical model selection, it is customary to use a model selection criterion C as the decision procedure. Theoretically, we may assume that the criterion acts as a function over the values v_{jk}. The appropriateness of a criterion for an individual i can then be evaluated as the match between the weights which the criterion inherently assigns to

the values v_{jk} and the weights w_{ij} the decision-maker i assigns to the aims g_j.

Because disagreement on the methodological level must be resolvable by reaching agreement on the axiological level, the hierarchical theory is based on the assumption that a single stance on the axiological level unequivocally leads to a single stance on the methodological level. Thus, there should not be multiple conflicting methods that are optimal in the light of the values by which the methods are judged. Similarly, the hierarchical theory prescribes a direct relation between an evaluation method and a model: there should not be multiple, conflicting models that are optimal in the light of the agreed-upon evaluation methods. Thus, the usefulness of the hierarchical theory of justification in the context of statistical modelling rests on the existence of an optimal model evaluation criterion C^* in a set $\mathcal{C}$ of proposed criteria, given a specific stance on the axiological level, and the existence of a model M^* in a set $\mathcal{M}$ of proposed models, that is optimal in the light of C^*. More formally, the two critical assumptions of the hierarchical model are that, for any (non-empty) set of aims $\mathcal{G}$, criteria $\mathcal{C}$, and models $\mathcal{M}$:

$$(\exists C_i \equiv C^*, C_i \in \mathcal{C})(\forall C_{j\neq i} \in \mathcal{C})(u_{|\mathcal{G}}(C_i) \succ u_{|\mathcal{G}}(C_j)), \tag{3.1}$$

in which $u_{|\mathcal{G}}$ denotes the utility of a particular criterion in forwarding a set of aims $\mathcal{G}$, and

$$(\exists M_i \equiv M^*, M_i \in \mathcal{M})(\forall M_{j\neq i} \in \mathcal{M})(C^*(M_i) \succ C^*(M_j)). \tag{3.2}$$

3.1.4 Overview

The purpose of this chapter is to determine the tenability of these two assumptions and with that of the hierarchical theory of justification. The field of statistical model selection is especially useful for this analysis. Due to the quantitative nature of statistical models, the epistemic consequences of different models (the values of models relative to epistemic aims) can be relatively precisely determined. The hierarchical theory is often assumed to hold, either explicitly in discussions of scientific rationality and theory choice, or implicitly in those of statistical model selection. For the latter field, the hierarchical theory may provide a general framework in which to address different model evaluation criteria. To my knowledge, statistical model selection has not been viewed as a multi-attribute decision problem. This view can provide a means of overarching the separation in the literature on model selection, couched in different statistical paradigms such as frequentism, Bayesianism, and information-theory. Even if the main assumptions are untenable, the quality of the hierarchical theory as a descriptive device may remain relatively untouched. The untenability of the hierarchical theory as a normative device does have implications for the view that model selection can proceed on an algorithmic basis. Furthermore, the tenability of the hierarchical theory has strong implications for the classical view that scientific consensus is the product of the (or at least, 'a') scientific method (e.g. Laudan, 1984).

The structure of the remainder of this chapter follows the structure of the hierarchical theory of justification. First, the epistemic aims comprising the axiological level are addressed. This is followed by a discussion of the model selection criteria and their relation to these epistemic aims. After this spade-work, the tenability of the two critical assumptions will be addressed. Essentially, this will require two forms

of determination, namely that the choice for a method is completely determined by a stance on the axiological level, and that the choice for a model is determined by a stance on the methodological level.

3.2 Axiological values: precision, generality and simplicity

An analysis of the applicability of the hierarchical theory of justification to the situation of statistical model selection naturally starts at the top level in the hierarchy: what specific values are taken into account when a statistical model is evaluated? From a strictly realist viewpoint, an optimal model is that which represents the 'true' relations in the population, so that $M^* = F(\boldsymbol{x})$. From a strictly instrumentalist viewpoint, an optimal model is that which is most accurate in predicting observations. A model that is optimal from a realist viewpoint will be optimal from an instrumentalist viewpoint, but this relation is not necessarily reciprocal. In practise, a model's predictive qualities may be the sole indication of a model's optimality available to proponents of either view. Two models performing equally well in this respect, such as the two polynomials in the curve-fitting example, must be distinguished on other criteria, based on other axiological values.

According to Forster (2000), all scientific approaches to model selection follow three steps: the specification of a goal, the specification of a criterion as means to this goal, and an explanation of how the criterion achieves the goal. This characterisation is similar to the hierarchical theory in that the goal is the axiological value, the criterion or means is the method and the explanation should state how the method exactly follows from the goal. As stated earlier, the goal for a realist would be to arrive at the true model, while the goal for an instrumentalist would be to arrive at the best predicting model. The problem for the realists is how to reach the goal of a true model, since there is no universally valid method to determine the truth of a model. The task for the instrumentalist may be easier, although a method for proving that a model will optimally predict future observations is also not at hand. Since the axiological values of truth and optimal (perfect) prediction do not directly lead to methods for their realisation, the hierarchical theory may be rejected. However, this is premature if other aims can be distinguished which are directly associated with methods for their realisation.

The axiological values considered relevant in the context of statistical modelling are precision, generality and simplicity. Although still other values may enter discussions regarding the appropriateness of models, these three values seem commonly attached to statistical models and theories in general. For instance, Popper (1959) summarises the requirements of good theories – theories high in empirical content – as precision, universality (generality) and low dimensionality (essentially 'simplicity') In a recent special issue of Journal of Mathematical Psychology (vol. 44, issue 1) on statistical model selection, most authors base their description on some or all of these values. Jacobs' and Grainger's (1994) list contains descriptive adequacy, generality, simplicity (and falsifiability) and explanatory adequacy, in which the last one is described rather vaguely. Finally, Myung and Pitt (1997) name descriptive adequacy and complexity, although their definition of complexity seems to incorporate the generality of a model.

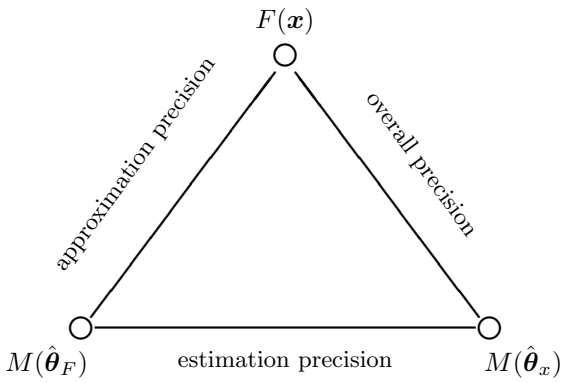


Figure 3.2: Types of precision as relations between $F(\boldsymbol{x})$, $M(\hat{\boldsymbol{\theta}}_F)$ and $M(\hat{\boldsymbol{\theta}}_x)$

In the following paragraphs, precision, generality and simplicity are first treated separately. Since in practice, they are often interrelated or overlapping, some of their connections are described in the closing section of this paragraph.

3.2.1 Precision

Of the three values, precision is least problematic (Cutting, 2000). Using the framework of Linhart and Zucchini (1986), we may make a distinction between a model's approximation precision and its estimation precision (see Figure 3.2). A model's approximation precision is its maximum obtainable precision in representing $F(\boldsymbol{x})$. It is the inverse of the discrepancy $\Delta[F(\boldsymbol{x}), M(\hat{\boldsymbol{\theta}}_F)]$ between the data generating process and the best approximating model. For a correctly specified model, the approximation precision clearly equals its maximum value. A model's estimation precision is the precision of the fitted model in representing the best approximating model. It is the inverse of the discrepancy $\Delta[M(\hat{\boldsymbol{\theta}}_x), M(\hat{\boldsymbol{\theta}}_F)]$ between the fitted and best approximating model. The estimation precision thus represents a fitted model's precision in the light of sampling error and is dependent on sample size. With the types of precision distinguished here, the realist aim can now be defined as maximising approximation precision, while the instrumentalist aim is defined as maximising the expected overall precision, which effectively consists of finding an optimal balance between approximation and estimation precision. Since the data generating process is generally unknown, it is not straightforward to show the extent in which either aim is met, although the instrumentalist aim may be the more realistic one.

Maximal precision per se is usually not the objective in statistical modelling (e.g. Kaplan, 2000). The specification of a model should not be motivated by the need for a complete description of reality; if this would be the aim, then why specify a model at all? Reality is a precise enough description of itself; no model could do better! A model is taken to be an analogy for reality and should highlight those particular aspects of it that are interesting in the context of what one wants to explain.

The precision of a model should be adequate, not necessarily complete. Thus, for a statistical model, precision becomes descriptive adequacy (e.g. Jacobs & Grainger, 1994). Taking observations as the sum of signal and noise, adequate precision would be reached when the model addresses the signal and not the noise. Of course, whether the residual variation is really random noise is a metaphysical question, whose answer is usually imposed by assumption.

3.2.2 Generality

Generality is more difficult to specify than precision (Cutting, 2000). It can be defined in terms of the domain to which the model applies: a model built in the domain of one data set hopefully generalises to other data sets, observed in similar situations. In this sense, generality is defined as the predictive precision of a model. A model's maximum predictive precision equals the estimation precision defined above. A related but different aspect of generality is a model's scope, which can be defined as a model's robustness to changes in the data generating process. This term could be operationalised as the proportion of the model-appropriate domain in which a model is expected to reach a certain level of precision (Cutting, 2000; Popper, 1959). As such, it requires specification of the model-appropriate domain, or the set of data generating processes to which the model is required to apply. This aspect of a model may be studied by simulation. While it is desirable for a fitted model to perform well under certain changes in the data generating process, a model that is precise in all possible samples to which it may be fitted is regarded as too flexible and unfalsifiable. This is related to Popper's (1959) notion of the empirical content of a hypothesis: the more the predictive domain of a hypothesis is restricted, the more falsifiable it is and the more specific the information it entails. Analysis of a model's scope is obviously more involved, and usually only predictive precision is taken into account.

3.2.3 Simplicity

Simplicity, while arguably one of the most widely underwritten values in science, may be the hardest of the three values to define (Cutting, 2000; Popper, 1959). In contrast to precision and generality, which are properties of the relation between a model and the data, simplicity is a property of the model alone. Despite the lack of an universally accepted definition (Derkse, 1993), the principle of simplicity has guided scientific inference for over two millennia. In the Physics, Aristotle explicitly stated that it is better to use a minimal plurality of principles (see Derkse, 1993, for an analysis of Aristotle's use of simplicity). Currently, the principle of simplicity is commonly referred to as Occam's Razor, after William of Ockham, who applied the principle so often and with such rigour, that the principle of parsimony was later given his name. In the form mistakenly ascribed to Ockham (Thorburn, 1918), the principle is stated as 'plurality should not be posited without necessity', but in practice, the principle is usually taken to mean something like 'among the several theories that are all consistent with the observed phenomena, one should pick the simplest theory' (e.g. Li & Vitáni, 1992). But when is a theory or model simpler and why should it be preferred when it is? While the last question doesn't require an answer if simplicity

is purely a goal, it has been posed so often that it will be insightful to deal with it. For the attempts at justification lead to the question whether simplicity is held as a goal in itself or is assumed to be a principle which advances another goal.

How is simplicity defined?

For Aristotle and Ockham, the principle of simplicity referred mostly to paucity in the presupposition of entities in the world. For instance, Ockham used the principle to defer with the existence of species, the medieval concept of mediators between object and knower. The assumed simplicity of objects in reality can be termed ontological simplicity, to distinguish it from semiotic simplicity, which refers to the simplicity of concepts, propositions and theories (Bunge, 1962). Ontological simplicity is mostly associated with ontological justifications, while more diverse justifications are associated with semiotic simplicity. The latter type of simplicity is usually assumed in current versions of Occam's Razor, and it will be this form of simplicity that is dealt with here. Moreover, as Bunge (1962) argues, this is the primary form of simplicity, since ontological simplicity can be judged through scientific analysis. One problem of semiotic simplicity is its possible language dependence. In terms of the curve-fitting example, a curve that looks simple in a $x - y$ diagram will look complex in a $x' - y$ diagram, when x' is a nonlinear transformation of the original variable x. To compare the simplicity of different theories, one would like a criterion independent of the language in which the theory is stated. One solution to this problem has been to define the simplicity of an object x in terms of its Kolmogorov complexity, which is defined as the shortest binary computer program that has x as output (Li & Vitáni, 1992). Since the length of a program depends on the language in which it is written, the definition requires a universal computer language, defined in terms of a Universal Turing Machine, capable of describing all other Turing machines. By referring to the shortest binary program for a Universal Turing Machine, the Kolmogorov complexity provides a language invariant definition of simplicity. Unfortunately, it is logically impossible to compute. However, different estimations of the Kolmogorov Complexity have been proposed, such as Rissanen's (1983) Minimum Description Length, which will be discussed later.

For statistical models concerned with relations in a finite set of observations, language variance may not pose a direct problem, since the models considered can be formulated in terms of probability distributions over the outcome space. The simplicity of a statistical model is then usually defined in terms of the number of adjustable model parameters, an approach similar to that proposed by Jeffreys (cited in Sober, 2000). According to Jeffreys, the simplicity of a law is defined as the summation of the number of freely adjustable parameters with the absolute values of its integers (degrees and datives). However, this definition is difficult to apply to a function such as $y = \sin x$ (Sober, 2000). Similarly, Popper (1959) defines the simplicity of a theory in terms of the number of its dimensions, which, for quantitative laws equals the number of freely adjustable parameters.

Although the simplicity of a model and the number of parameters will be generally related, the strength of this relation is not always clear. As the above example makes clear, the 'number of free parameters' measure of complexity may not capture

all intuitively clear simplicity differences. Myung and Pitt (1997) propose a measure of complexity that, beside the number of model parameters, also takes the 'functional form' of a model and the 'extension of the parameter space' into account. The functional form of a model is defined as 'the way in which parameters are combined in the model equation' (p.81) and the extension of the parameter space is defined as the range of parameter values allowed by the model. Although the functional form seems incorporated in intuitive notions of simplicity, the extension of parameter space seems more characteristic for a model's flexibility or its precision in making predictions about parameter values. The example provided by Myung and Pitt is that of two models that are both characterised by the equation $y = 1/(1 + e^{-\theta x})$, but with different parameter spaces, $\mathbf{\Theta}_{M_1} = \mathbb{R}$ and $\mathbf{\Theta}_{M_2} = \mathbb{R}^+$. Essentially, M_2 is nested in M_1, the model M_2 being restricted to increasing functions, while M_1 also incorporates decreasing functions. This renders M_1 a more general model, but is it really a more complex one? Take for example two models characterised by the function $y = a + bx$, in which for model $M_1 : a \in (0, 2)$ and for model $M_2 : a \in (0, 1)$. Is model m_1 to be judged (more) 'complex', or less specific, but more general?

The concept of simplicity is complex (Bunge, 1962), and as yet there may not be a measure available that captures all of its aspects. Even for the relatively easy case of mathematically specified models, its true nature may remain elusive. However, there seems to be a large consensus in the statistical modelling field on the operationalisation of the simplicity as the number of model parameters (Cutting, 2000), and that is all that the hierarchical theory requires.

How is the principle of simplicity justified?

Despite the murky definition of simplicity, Occam's Razor is often taken as the canon of scientific inference. The classical motivation for applying the simplicity principle is ontological: a simple theory is to be preferred because this reflects the inherent simplicity of nature. This position was taken by Aristotle, who argued that if nature had unlimited constituent parts, it would be indefinite, in which case things of nature could not be the object of knowledge, a conclusion he deemed absurd (Derkse, 1993). For Ockham, the ontological motivation was also important, although he seldom explicitly justified his use of the simplicity principle. Also, Newton took an ontological stance when he stated that 'Nature is pleased with simplicity, and affects not the pomp of superfluous causes' (Principia Mathematica, 1687; cited in Thorburn, 1918). In the light of current theories, in which natural systems move towards complexity (Nicolis & Prigogine, 1989), such a stance seems difficult to defend.

For Popper (1959), the simplicity principle is justified by empirical content: 'Simple statements, if knowledge is our object, are to be prized more highly than less simple ones because they tell us more; because their empirical content is greater; and because they are better testable' (p.142). Popper illustrates his position with an example of the simple hypothesis of a straight line and the complex hypothesis of a circle: three data points are required to falsify the hypothesis that all points lie on a straight line, while four data points are required to falsify the hypothesis that all points lie on a circle. Similarly, Sober (1975) related the simplicity of a hypothesis to the amount of information it conveys. More specifically, he defines the simplicity of a

hypothesis in terms of the amount of extra information needed to answer a question related to the hypothesis.

Mach (1956, cited in Derkse, 1993) justifies the simplicity principle by psychological economy. According to Mach, science is defined as reducing the need for experience by the anticipation of facts in thought. Simpler theories are more cost-effective in this reduction. Mach's justification by psychological economy seems to follow from an instrumentalist view, since the psychological economy of a hypothesis in no way guarantees a model's truth. As long as models have a similar predictive precision, a strict instrumentalist, judging models purely on their predictive qualities, should be indifferent. This indifference leaves room for additional evaluations, so that the simpler may be preferred for its psychological economy, but this preference does not follow from a strict instrumentalist position. As Planck argued (cited in Derkse, 1993), every theory can be simultaneously a simplification and a multiplication, depending on one's point of view. For example, the hypothesis of atomism could be judged a legitimate theoretical simplification as well as a needless multiplication of entities. Basically, Mach's form of simplicity is not an inherent quality of an object, but a subjectively assigned label by a perceiver of the object. Each perceiver can assign a degree of simplicity to the object he or she sees fit, making the justification of the principle also subjective. This subjectivism makes Mach's version of simplicity unfit to function as an axiological value in the hierarchical theory of justification, since it does not lead to a single method for its realisation.

Jeffreys (1961) argues that simple models should be preferred because they are a priori more probable. He bases this assertion on a similar ground as Mach: scientists seem to prefer simple models to complex ones, all other things being equal. Considering nested models, the notion that simple models have a higher prior probability is rather counterintuitive. Standard probability theory prescribes that, if $\mathcal{X}$ is a subset of $\mathcal{Y}$, then the probability of $\mathcal{X}$ cannot exceed the probability of $\mathcal{Y}$, i.e. $\mathcal{X} \subset \mathcal{Y} \implies P(\mathcal{X}) \leq P(\mathcal{Y})$. So actually, a more complex model should have a higher probability than a simpler one, at least when the models are nested.

The final justification considered here is statistical in nature and relates simplicity to predictive accuracy by arguing that the additional parameters in a more complex model may just capitalise on chance, accounting for random sample fluctuations rather than true population characteristics. This is the traditional justification for trading off fit for the simplicity of a model (e.g. Cheeseman, 1990; Forster, 2000; Myung, 2000). Reichenbach (1938) has argued along this line in relation to the curve-fitting problem. As Zucchini (2000) points out, a more complex model M_1 will generally have a smaller error of approximation than a simpler model M_2, but the fitted model $M_1(\hat{\boldsymbol{\theta}}_x)$ will generally have a greater error of estimation than a fitted model $M_2(\hat{\boldsymbol{\theta}}_x)$. This is due to the general observation that, for a fixed number of observations, the parameter estimates for model M_1 will be relatively less stable than those for model M_2. Thus, while the more complex model family is potentially more precise, a realised model $M_1(\hat{\boldsymbol{\theta}}_x)$ from this family will generally not live up to its potential. This argument applies directly mainly in situations where model M_2 is nested in M_1, so that greater complexity is the result of additional model-parameters. In the presence of random sample error, a model that fits the data perfectly well accounts for the sample error

as if it were a structural aspect of the data generating process. Remember that from the definition of a correctly specified model, it follows that if a model M is correctly specified, those models in which M is nested are also correctly specified. Consider a sequence of nested models $(M_1, M_2, \ldots, M_k, M_{k+1}, \ldots)$, in which model M_k is correctly specified (and with that also models $M_{k+1}, M_{k+2}, \ldots$), while M_{k-1} isn't. We may call M_k the *minimal correctly specified* model. If there was an infinite amount of data, all correctly specified estimated models $M_k(\hat{\boldsymbol{\theta}})$, $M_{k+1}(\hat{\boldsymbol{\theta}}), \ldots$ would be practically identical (the additional parameters of the models $M_{k+1}(\hat{\boldsymbol{\theta}}), M_{k+2}(\hat{\boldsymbol{\theta}}), \ldots$ would for instance be estimated at 0). But since the data is limited, parameter estimates are unreliable, and will be influenced by the random sample error. And the more additional parameters to estimate, the larger such influence. In short, the argument for simplicity rests on the more complex models $M_{k+1}, M_{k+2}, \ldots$ having additional parameters which are *redundant* in the light of predictive accuracy, since they only account for random error in a particular sample. But since non-nested models may differ in respect to their approximation precision, so that a more complex model is correctly specified, while a simpler is not, the argument of greater predictive accuracy does not necessarily apply in this case.

From all justifications of the simplicity principle, the last may be most appropriate for the selection of statistical models. The trade-off between model fit and simplicity to maximise predictive accuracy seems to constitute a well-defined version of Occam's Razor (Forster, 2000). However, since this justification applies directly only when choosing from a set of nested models, this version of Occam's Razor may be restated as: For a set of nested models, the preferred model is that which does not fit the data significantly worse than the other models, while its parameter estimates are more reliable than those of the others. This seems justifiable from an instrumentalist viewpoint and it makes the model selection partly dependent on the sample size. From a realist viewpoint, however, this justification does not make immediate sense, since a model should be judged on its correspondence to reality, not the precision with which the parameters of a possibly misspecified model can be estimated in a limited sample of observations.

3.2.4 Connections

A link between the simplicity of a model and its predictive accuracy was described above as a justification for the simplicity principle. This is based on the simple assertion that, in general, for a fixed number of observations, the reliability of parameter estimates is inversely related to the number of model parameters. Thus, simplicity defined in terms of the number of model parameters may be positively related to a model's predictive accuracy. More precisely, for two nested, correctly specified models, the estimated simpler model will more likely reliably represent the structural relations in the data generating process, since the more complex will more likely incorporate random sample fluctuation into its parameter estimates. Similarly, in the presence of random sample fluctuation, the descriptive precision and predictive accuracy are also inversely related. Finally, in the presence of sample error, there will be a negative relation between the simplicity of a model and its descriptive accuracy. Note that these relations directly hold for nested, correctly specified models. The argument for

all relations is similar: a model should account for important relations in a population and these relations may be more reliably estimated in simple models. However, for misspecified models, the parameter estimates of a simpler model may be more reliable, but less valid.

3.3 On the methodological level: model evaluation criteria

The dependencies between the axiological values show that they cannot all be optimally realised at once. There is often a trade-off between simplicity and precision and between precision and generality. A model evaluation criterion should decide on the model in which this trade-off is optimally reached. According to the hierarchical theory, the specific model evaluation criterion to use should follow from one's stance on the axiological level. That disagreement concerning the appropriate model evaluation criteria is likely, follows from the number of available fit-indices, a number so large that one author remarked that 'it seems likely we will need a tool to select model selection tools in the not too distant future' (De Leeuw, 1990, p. 240).

In the following sections, common model evaluation criteria are compared by identifying the axiological values they address. This is not a straightforward matter, since a number of indices simultaneously address multiple values. However, one value may be the dominant basis in a given criterion, and the following categorisation rests on such dominance.

3.3.1 Criteria addressing precision

A model's descriptive precision is directly observable, and a model evaluation criterion that incorporates it directly can be based on the χ^2 discrepancy

$$C_\chi = \sum_{i=1}^{n} \frac{\left[f(\boldsymbol{x}_i) - M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})\right]^2}{M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})}, \tag{3.3}$$

or the Gauss discrepancy

$$C_\mathrm{G} = \sum_{i=1}^{n} [f(\boldsymbol{x}_i) - M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})]^2. \tag{3.4}$$

Another criterion can be defined in terms of the log-likelihood

$$C_\mathrm{L} = -\tfrac{1}{n} \sum_{i=1}^{n} \log [M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})]. \tag{3.5}$$

In the absence of sampling error, such criteria would be good indicators of the approximation precision, but if the observed distribution is a random realisation of the data generating process, minimising these discrepancies will probably lead to overfitting, and a researcher should then be more interested in a model's average precision over a number of such realisations, i.e. the predictive precision.

3.3.2 Criteria addressing generality

A natural definition of predictive precision is in terms of the Kullback-Leibler distance

$$\Delta_{\mathrm{KL}}[F(\boldsymbol{x}), M(\hat{\boldsymbol{\theta}})] \tag{3.6}$$

between a model and the data generating process. A model evaluation criterion that provides an asymptotic estimate of the mean Kulback-Leibler distance of a model is the Akaike Information Criterion (Akaike, 1992[1973]), defined as

$$C_{\mathrm{AIC}} = -2\sum_{i=1}^{n} \log\left[M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_x)\right] + 2k, \tag{3.7}$$

where k is the number of free model parameters. C_{AIC} is sometimes interpreted as a criterion of predictive precision (e.g. De Leeuw, 1992) , and sometimes as a criterion that penalises a model's descriptive precision with its complexity (Forster & Sober, 1994). A criterion with a similar form, although derived in a different theoretical framework, is the Bayesian Information Criterion (Schwarz, 1978), defined as

$$C_{\mathrm{BIC}} = -2\sum_{i=1}^{n} \log\left[M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_x)\right] + k\log n. \tag{3.8}$$

A model for which C_{AIC} or C_{BIC} is minimal should be chosen. The difference between C_{AIC} and C_{BIC} is that the first selects the model of which the estimated Kullback-Leibler distance is minimal, while the latter selects the model which maximises the posterior probability of the model given the data (Myung, 2000).

3.3.3 Criteria addressing simplicity

As previously mentioned, C_{AIC} and C_{BIC} can be interpreted as criteria that penalise a model's fit by its complexity. The penalty term $2k$ of C_{AIC} is a constant for a given model, while the penalty term $k\log n$ of C_{BIC} is dependent on sample size, so that with increasing sample size the penalty for model complexity becomes relatively more severe. A criterion that specifically addresses a model's simplicity is the Minimum Description Length (Rissanen, 1996), which estimates the Kolmogorov Complexity by replacing algorithmic complexity with stochastic complexity (the shortest obtainable description of $\boldsymbol{x}$ by a model class M). The MDL takes the familiar form of a penalised likelihood and is defined as

$$C_{\mathrm{MDL}} = -\sum_{i=1}^{n} \log\left[M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_x)\right] + \tfrac{k}{2}\log\tfrac{n}{2\pi} + \log\int\sqrt{|I(\boldsymbol{\theta})|}d\boldsymbol{\theta} + o(1) \tag{3.9}$$

(Rissanen, 1996), in which $|I(\boldsymbol{\theta})|$ is the determinant of the Fisher information matrix and $o(1)$ becomes negligible for n large. The last terms are often difficult or impossible to compute, but a reasonable practical version views stochastic complexity as a two-stage description of the data, consisting of the encoding of a model $M(\boldsymbol{\theta})$ and the encoding of the data $\boldsymbol{x}$ using $M(\boldsymbol{\theta})$. This leads to an approximation of C_{MDL} as

$$C_{\mathrm{MDL}} = -\sum_{i=1}^{n} \log\left[M(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_x)\right] + \tfrac{k}{2}\log n \tag{3.10}$$

(Grünwald, 2000), which is identical to one half of C_{BIC}. As such, both C_{MDL} and C_{BIC} will result in an identical ordering of the models $M_i \in \mathcal{M}$.

3.3.4 Connections

The general form of a model evaluation criterion is that of a penalised likelihood, in which the penalty is dependent on the number of model parameters (C_{AIC}), combined with the number of observations (C_{BIC}, C_{MDL}). For a set of nested, correctly specified models, these criteria essentially compare the models on their expected predictive accuracy. But, given the prevalent view that statistical models are never true (e.g. McDonald & Marsh, 1990), it is questionable whether a model is ever correctly specified. As Golden (2000) has shown, when dealing with two non-nested models, one of which is correctly specified and the other not, the model selection criteria may actually show a strong favour for the misspecified model. Golden proposes a model selection test, which tests the hypothesis that two models have an equal error of approximation. This test is highly conservative, leading to a choice in just a few situations. At least for the applicability of the hierarchical theory of justification, a method that mostly results in indifference is not an optimal method.

3.4 The hierarchical theory in action

As indicated earlier, the applicability of the hierarchical theory as a normative device rests on the two critical assumptions given in (3.1) and (3.2). Below, the tenability of these two assumptions will be discussed in turn.

3.4.1 From aims to methods

The first critical assumption (3.1) of the hierarchical theory of justification is that by taking a stance on the axiological level, the number of optimal evaluation criteria is restricted to one. Since precision, generality, and simplicity are all generally endorsed and multiple criteria address all three aims, this restriction either depends on (a) a single method optimally addressing all values, or (b) a correspondence between an individual's relative ranking of the aims and the relative strength with which the criteria address them.

Since the optimality of the evaluation criteria depends on the correctness of the assumptions on which the criteria are based, which in turn requires knowledge of the data generating process, I take option (a) not to be true for realistic situations. The problem here is that both the adequacy of a criterion and a model rest on assumptions regarding the data generating process $F(\boldsymbol{x})$. For instance, C_{BIC} is consistent, meaning that if $F(\boldsymbol{x}) \in \mathcal{M}$, then by using C_{BIC}, $P[M^* = F(\boldsymbol{x})] \to 1$ when $n \to \infty$. This does not hold for C_{AIC}. On the other hand, C_{AIC} is loss-efficient when $F(\boldsymbol{x}) \notin \mathcal{M}$, meaning that the expected squared error of M^* is asymptotically equivalent to the smallest possible of all candidate models in $\mathcal{M}$. This does not hold for C_{BIC}. The assumption that at least one of the models in $\mathcal{M}$ is correctly specified cannot be tested. Since the set of models $\mathcal{M}$ will not exhaust the possible models, it is neither always true. Option (b) seems more viable. For instance, since C_{MDL} and C_{BIC} place a stronger penalty on

extra parameters as the sample size increases, researchers putting more emphasis on simplicity than generality may prefer these criteria to C_{AIC}, while researchers putting more emphasis on generality may choose the latter.

3.4.2 Model equivalence: statistical underdetermination

The second critical assumption (3.2) is that the choice for a model evaluation criterion restricts the number of optimal models to one. Regarding this requirement, the problem of underdetermination must be addressed. As exemplified in the introduction of this chapter, multiple models may fit the data equally well, so that their choice is underdetermined by precision. Such equivalent models arise routinely, especially in the area of structural equation models, which is one of the more prevalent types statistical model in psychology. In structural equation modelling, two models M_1 and M_2 are equivalent if they reproduce the same set of covariance matrices when their parameters vary over the parameter spaces. The topic of model equivalence in structural equation modelling has received some attention over the last 20 years (e.g. MacCallum, Wegener, Uchino, & Fabrigar, 1993; Raykov & Penev, 1999; Stelzl, 1986; Williams, Bozdogan, & Aiman-Smith, 1996), leading to various algorithms for producing equivalent models. One of the more sophisticated algorithms is based on graph theory, and allows the specification of the entire equivalence class for a particular model M. Although the problem of model equivalence has received due theoretical attention, in practice, scientist rarely acknowledge the existence of equivalent models for a favoured model (Breckler, 1990). As Markus (2002) argues, equivalent models pose a problem only to the extent that the statistically equivalent models are not semantically equivalent. Trivially, any model is equivalent to itself. Also trivially, any model is equivalent to a nesting model in which all additional parameters are set to 0 (or 1). Such examples of equivalent models are not very interesting, because these models will have an identical semantic interpretation. Examples of equivalent structural equation models that do differ semantically can be easily constructed by changing the causal direction of certain effects. Such examples may be seen as instances of the classic methodological adagio that 'correlation does not imply causation'. But this is not all there is to statistical model equivalence. An interesting example of model equivalence is given by Bartholomew and Knott (1999), who show that a $(k+1)$-latent profile model is equivalent to a k-factor model. Molenaar and Von Eye (1994) expanded on this result, showing the complete equivalence of these models on the level of second-order moments. These authors pointed to the significance of this result for the types vs traits discussion in differential psychology. Essentially, if latent profile analysis was more popular than factor analysis, we might be talking about the Big Six rather than the Big Five.

Equivalent models pose a serious problem to the hierarchical theory of justification. Two types of statistical equivalence can be distinguished. Sample equivalence means that for a given finite sample of data $\boldsymbol{x}$, two models M_1 and M_2 have identical likelihoods, i.e.

$$\sum_{i=1}^{n} \log\left[M_1(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_x)\right] = \sum_{i=1}^{n} \log\left[M_2(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_x)\right]. \tag{3.11}$$

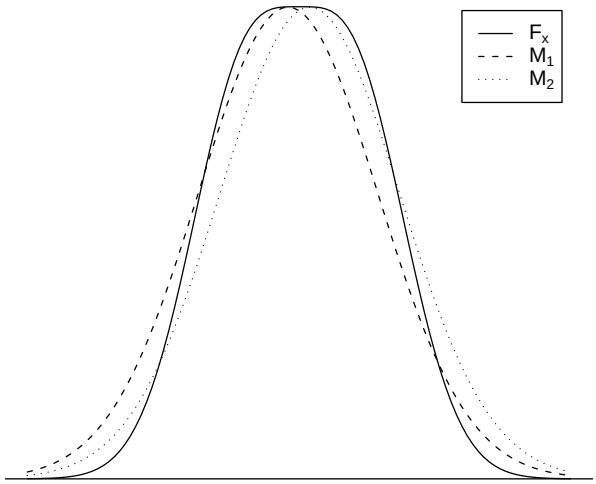


Figure 3.3: Population equivalent models

Population equivalence means that models assign identical likelihood to the population of all data generated by the data generating process $F(\boldsymbol{x})$. This implies that two models M_1 and M_2 are population equivalent if

$$\Delta_{\mathrm{KL}}[F(\boldsymbol{x}), M_1] = \Delta_{\mathrm{KL}}[F(\boldsymbol{x}), M_2]. \tag{3.12}$$

Population equivalent models may have identical likelihood functions, as is the case with many equivalent SEM models. But this is not necessarily so. An example of two population equivalent models with different likelihood functions is given in Figure 3.3. Note that models that are population equivalent are not necessarily sample equivalent, since population equivalent models can assign a different likelihood to a given sample. But as sample size increases, so does the probability that two population equivalent models will also be sample equivalent. Population equivalence results in the more serious form of underdetermination, since population equivalent models are underdetermined by all possible observations. Sample equivalence is less serious, since gathering more data will usually resolve the underdetermination.

Since equivalent models assign identical likelihood to observed data $\boldsymbol{x}$, they must be distinguished by other aspects than descriptive precision. The two aims endorsed most widely are generality and simplicity. The underdetermination problem is not solved by a generality criterion defined in terms of predictive precision, since equivalent models give similar predictions for new data. If equivalent models have different numbers of parameters, they can be distinguished by a simplicity criterion. However, equivalent models will often have the same number of parameters (Williams et al., 1996), so that none of the model selection criteria described here will be able to distinguish between them.

3.5 Conclusion

The problem of statistical modelling has been described as the choice for an optimal model M^* from a set of possibly misspecified, possibly non-nested models $\mathcal{M}$. According to the hierarchical theory of justification, disagreement concerning this choice is resolved by reaching agreement on the proper model evaluation criteria. Disagreement concerning these model evaluation criteria in turn is resolved by reaching agreement on the axiological values, defining the proper goals for statistical modelling or science in general. The hierarchical theory defines a top-down mechanism where agreement on the axiological level forges consensus on the proper model evaluation criteria, which in turn forges consensus on the proper model M^*. It is a normative theory, describing a rational method of resolving scientific disagreement. The applicability of the hierarchical theory to statistical model selection is limited due to two reasons. The first is that the optimality of model selection criteria rests on assumptions regarding the data generating process $F(\boldsymbol{x})$. This means that the problem of deciding on an optimal selection criterion can not be entirely separated from that of deciding on the optimal model. The second is the possible existence of equivalent models. This means that two models may be indistinguishable on the basis of a chosen model selection criterion. Admittedly, this conclusion is based on a specific set of axiological values and model evaluation criteria. Incorporating other axiological values, or different operationalisations thereof, may lead to different model evaluation criteria which do distinguish between otherwise equivalent models. For example, some authors argue that the definition of simplicity in terms of the number of parameters is not adequate (e.g. Bozdogan, 2000) and that the functional form of the model should also be taken into account. In the ICOMP(IFIM) criterion (e.g. Bozdogan, 2000), complexity is defined in terms of the inverse of the estimated Fisher Information matrix, of which the elements represent the (co-)variances of parameter estimates. While this criterion, as well as the proper version of C_{MDL}, might be able to distinguish between otherwise equivalent models, it is interesting to note that here it is not the simplicity of a model that seems to matter, but the reliability of parameter estimates. This is also the main concern when using C_{AIC}, which amounts to choosing the model with the highest expected overall precision. Since the estimation precision will increase with sample size, it makes sense that the effect of the penalty term disappears with increasing sample size. As such, it can be argued that the simplicity aim is parasitic on the aim for reliable parameter estimates. This holds to a lesser extent for C_{BIC} and C_{MDL}, where the penalty on extra parameters becomes relatively stronger when sample size increases. As such, these criteria may be the more appropriate when simplicity per se is an aspired aim. However, it is unclear whether simplicity can be taken to serve the higher-order realist or instrumentalist aims. Taken as the maximisation of approximation precision, the realist aim is served by simplicity only under the ontological assumption that nature (the data generating process) is inherently simple. When the instrumentalist aim is defined as the maximisation of expected overall precision, the choice between two equivalent models performing equally well in this respect should be considered arbitrary. The main justification for the simplicity principle in statistical modelling is related to the reliability of parameter estimates: a simple model that describes the data sufficiently well will have a better predictive accuracy because its

parameter estimates are less dependent on random sample fluctuations. When there are no other reasons to prefer simple models, why not simply prefer reliable models? This position would make sense, since it states that no more assertions are to be made than those that are justified by the data at hand. Moreover, the concept of reliability is well defined in statistics, while simplicity is not well defined in or outside of statistics. Although the simplicity principle and the reliability principle may appear identical, they are not. Simplicity is an aspect of the model itself, while reliability concerns a relation between the model and the observed data.

Since, at least when dealing with equivalent models, theoretical considerations must enter the model evaluation, the hierarchical theory paints an incomplete picture of justification. Disagreement on the theoretical level cannot be unequivocally settled by agreement on higher levels. In a sense, model selection is doubly underdetermined: methods are underdetermined by aims and theories are underdetermined by methods. In this regard, the 'battle on theory' must be fought at least partially on home territory.

4
Underdetermination and social validation in inductive tasks

In inductive tasks, the objective is to discover the process that generated a given body of data. Such tasks are difficult due to problems of underdetermination (see Chapter 2). On the basis of the general underdetermination thesis it has been argued that since theory choice cannot be sufficiently based on empirical or logical criteria, it should be explained by social factors (Kuhn, 1970; Hesse, 1980). There is a clear connection between this idea and one of the cornerstones of social psychology. According to the theory of social comparison (Festinger, 1954), when objective means to ascertain the correctness of a belief are lacking, people will tend to evaluate their opinions by comparing them to those of others. When a belief is underdetermined by empirical data, such objective means are lacking. In such circumstances, Festinger (1950) proposed that the validity of belief is established in social reality, rather than physical reality. In this social reality, consensus is the measure of valid belief. Social comparison tends to result in uniformity of opinion, for someone who disagrees with others in a group will tend to change his or her belief in the direction of the group position. Since its invocation by Festinger, the scope of social comparison theory has become increasingly wide, including the comparison of ability, belief, values and emotions. There are important differences between these domains. For instance, when ability refers not to the question 'Can I do X?', but to the question 'How good am I at X?', ability becomes a social construct. Since there is no objective standard of 'good' to compare one's ability to, it is a question of one's rank in a group and requires social comparison by definition. Beliefs, on the other hand, refer to potentially verifiable assertions about the true nature of an entity (Jones & Gerard, 1967). To answer the question 'Is belief X correct?', social comparison is neither necessary nor sufficient, but it can be informative. There are two main reasons why others' beliefs are informative. The first

is that different individuals may have based their beliefs on different informational bases. The second is that different individuals may process the same information in different ways. Different people may have different 'mental models' of a situation and some models may be more adequate than others. Inadequate mental models, those that do not fit the objective world very well, will lead to erroneous processing of the information and hence incorrect belief. Both lack of information and erroneous processing of information may lead to biased belief and when it is impossible to directly compare a belief to an objective standard the only means of investigating this bias may be through social comparison. The goal of correcting the two types of bias should result in the preference for dissimilar comparison others. To investigate bias stemming from insufficient information, one should compare one's belief with those held by people with different or more information. To counter bias from erroneous processing of information, one should compare one's belief with those held by people with the same information but different mental models. This is in disagreement with Festinger (1954), who proposed that people prefer to compare their opinions to those of people whose opinions are not too diverging from their own. This similarity thesis has been a controversial aspect of the theory. When taken literally, it defines social comparison as a conservative process directed at the preservation of existing opinion (Earle, 1986). This is clearly at odds with the central objective of social comparison: uncertainty reduction. By basing the selection of comparison others on agreement, the outcome of the social comparison is given beforehand and so the observed agreement has no informational value. The preference for dissimilar others for belief comparison was proposed in Goethals' and Darley's (1977) attributional reformulation of social comparison theory. Founded on Kelly's attribution theory, they proposed that social validation of belief concerns assessing whether a belief is entity or person caused. Beliefs that are entity caused are taken to be proper reflections of the entity to which the belief refers. Person caused beliefs, on the other hand, reflect idiosyncratic aspects of the person rather than an objective entity. In this account, agreement with similar others is uninformative, since the agreement may simply be the result from shared bias. Agreement with dissimilar others provides a strong indication the belief was entity caused. The preference for dissimilar others has been supported by a number of experiments (Goethals, 1972; Goethals, Darley, & Kriss, 1978; Gorenflo & Crano, 1989). A further refinement of Goethals' and Darley's attributional theory by Suls, Martin, and Wheeler (2000) further distinguishes social comparison processes as they relate to preference, preference predictions, or belief. According to the triadic model of opinion comparison (Suls et al., 2000), people prefer dissimilar advantaged others for belief comparison. The relevant attribute on which comparison others are chosen is the amount of information on which their belief is (assumed to be) based. Although comparing with people with more information is clearly a reasonable method of investigating bias due to lack of information, the triadic model overlooks the possible bias due to erroneous processing of information. The belief of others, who have based their belief on exactly the same information should be considered highly informative for this goal.

Moving back into the realm of inductive tasks, social comparison concerns factual beliefs about the generating process of observed data. For such objective judgements – that is, potentially verifiable statements concerning a factual matter – one may

expect strong effects of social comparison (Olson, Ellis, & Zanna, 1983). Because such judgements should not be influenced by subjective values, it is reasonable to expect similarity in judgement, given that all have adequate information to base their judgement on. Beliefs are an especially interesting domain in which to study social comparison, because the entity to which the belief pertains is not itself a social construct. Social constructs such as ability are defined by interpersonal comparison, so that social comparison is a necessary means for assessing ability. Because beliefs are true or false irrespective of personal values, variation in belief indicates that at least one of the disagreeing parties is wrong. Social comparison can provide valuable information regarding the truth of belief, especially if one assumes that beliefs that receive more social support are more likely to be true. Belief comparison is a relatively understudied aspect in social comparison research, which has mainly focused on the comparison of ability or value-type opinions rather than factual beliefs. In the studies that address belief comparison, the focus has been mainly on the preference for different types of comparison others (Gorenflo & Crano, 1989; Suls et al., 2000), rather than the process and outcome of the actual comparison process. However, indirect evidence for the social comparison of belief can be found in related experiments on conformity. Social comparison processes have been proposed to explain the conformity found in Asch-type experiments (Allen & Wilder, 1977). In these studies, subjects adopted an unanimously endorsed group position even when this position was clearly at odds with empirical evidence (Asch, 1952; see Allen, 1965, and Levine & Thompson, 1996, for an overview). Some of these conformers reported a social comparison motive: they simply couldn't believe they were right and all others wrong (Asch, 1952, p. 470-471). In the typical conformity experiment, the task of comparing lines to a baseline is unambiguous, while social comparison is only expected under uncertainty. When available and sufficient, people should rely solely on objective evidence and not on the opinions of others. On the other hand, participants in the conformity experiments had no choice but to compare their opinion to those of others, since all others voiced their opinion first. Social comparison theory concerns voluntary evaluation, while the conformity research imposes a situation of forced comparison (Allen & Wilder, 1977). Orive (1988) proposed that people use social comparison implicitly, by assuming agreement with similar others. In general, no attempt is made to test the validity of this social projection through explicit social comparison. An imposed disconfirmation of assumed consensus, such as in the conformity experiments, will induce uncertainty about the correctness of the belief. When this uncertainty is high enough, action is taken to reduce it. Two possible actions are particularly relevant in the present context. The first is to render the other as irrelevant for the evaluation of belief, for instance, by attributing the disagreement to the lower ability of the disagreeing party. The second action is to minimise the disagreement by changing one's belief in the direction of the others' beliefs, resulting in conformity. The unambiguity of the task in the conformity experiments may have actually moved people to the latter action, since participants have no reason to suspect others to perform worse than themselves in such an easy task.

According to Deutsch and Gerard (1955), conformity can result from two types of group influence: normative influence, defined as 'influence to conform to the positive expectations of another', and informational influence, defined as 'influence to accept

information obtained from another as evidence about reality'. The latter type of influence corresponds with social comparison theory. From the many variations in the conformity paradigm (see Allen, 1965 for an overview), it appears that both types of influence play a role in conformity. Reducing the normative pressure by breaking the group consensus with a deviant besides the participant reduces conformity considerably. In support of informational influence, the magnitude of the reduction in conformity depends on the assumed ability of the social supporter. For instance, Allen and Levine (1971) found that the reduction in conformity was much greater when the support came from a valid as compared to an invalid (i.e. visually impaired) source. While the results of the conformity experiments can be taken to support the role of social comparison in the formation and change of belief, these results can also be interpreted as disconfirming Festinger's theory since people are supposed to rely on social comparison only when objective information is insufficient for belief evaluation. Moreover, the percentage of people for whom the conformity was accompanied by a change in belief may have been very small. Allen (1965) doubted whether there was any private acceptance of the group position in the Asch-type experiments at all. About twenty years ago, Insko, Drenan, Solomon, Smith, and Wade (1983) noted that the evidence for conformity as a result of informational influence was very sparse, and this situation has not changed much since. The present study attempts to fill this hiatus by investigating social comparison processes in inductive tasks.

Inductive tasks provide an especially interesting situation for social comparison processes because they are objective in principle, but can be ambiguous due to underdetermination problems. In natural situations, most beliefs will be based on a combination of objective and social information. In studies on social comparison, usually only consensus information is provided, so the results may generalise poorly to natural situations (Insko et al., 1983). By providing evidence pertaining to the generating process, a naturalistic setting arises in which belief can be based on both objective and social information. Laughlin has conducted many experiments on social processes in inductive tasks, resulting in his theory of collective induction (Laughlin, 1999). In these studies, real groups collaborated to discover rules that generated sequences of playing cards. The group processes were analysed by means of Social Decision Schemes (Davis, 1973), and the following process was deemed adequate to describe the combination of individual hypotheses into a group hypothesis: If at least two group members propose correct and/or plausible hypotheses, the group selects among those only. Otherwise the group selects among all proposed hypotheses. If a majority of members propose the same hypothesis, the group selects a hypothesis according to a majority rule. Otherwise, the group follows a proportionality process. The preference for hypotheses held by a majority, as well as the finding that the group chooses the correct hypothesis when proposed by at least two members, are not incompatible with social comparison theory. The objective in the SDS analysis, however, is to ascertain how groups combine individual hypotheses into a single group hypothesis. Although this is certainly interesting in its own right, it does not provide evidence for social comparison processes in the formation of belief. The groups were required to posit a single collective hypothesis and it is not clear whether the collective hypothesis corresponds to those held by the group members individually. The present study attempts to overcome this by studying individual induction in

a Crutchfield type setting where participants receive not only objective information, but also anonymous feedback regarding the hypotheses of other group members. This feedback is manipulated to result in consensus or dissensus among the other group members. When hypotheses regarding the generating process are underdetermined by the objective evidence, both a dissenter and a unanimous group of others may find their beliefs equally supported by the available objective evidence. In such a situation, social comparison theory predicts that the social support for the unanimous group hypothesis increases its subjective validity. If this increase is large enough, a dissenter should be persuaded to conform to the consensual position.

4.1 Experiment 1

Inductive tasks have both an intellective and a judgemental component (Laughlin, 1999). While some hypothetical generating processes can be shown to be inconsistent with the data (the intellective component), the choice between multiple hypothetical processes that are equally consistent with the data is a judgemental task. In the terminology used here, whether or not there is a judgemental component depends on whether the generating rule is underdetermined by the data. Social comparison is only expected for the judgemental aspect of the task, so that conformity is expected only when the generating process is underdetermined by the evidence. This hypothesis is tested by investigating conformity in a situation of underdetermination or determination. If conformity in belief is the result of informational influence, as it should be according to social comparison theory, then conformity is expected in the presence of underdetermination, but not determination.

The inductive task in this experiment is an adaptation of the one used by Laughlin and colleagues (Laughlin, 1999; Laughlin, Chandler, Shupe, Magley, & Hulbert, 1995), in which participants have to determine by which rule a sequence of playing cards is formed. Exemplars of the rule are presented sequentially and participants are asked to report their hypothesis of the rule after each new exemplar. The evidence is manipulated to ensure either underdetermination or determination of the rule by the evidence. After reporting their initial hypothesis, participants receive (bogus) feedback concerning the hypotheses of the other group members. This social feedback is manipulated to show either an emerging consensus among the group members or not. After receiving the social feedback, a new exemplar is added to the sequence, after which participants are again required to report a hypothesis. Intermitting the social feedback with new evidence before asking a new hypothesis allows participants to base their new hypothesis on a combination of objective and social evidence. For underdetermined rules, there are multiple hypotheses consistent with this new information of which the consensual hypothesis is one. Although the new evidence may be consistent with a participant's hypothesis, it is equally consistent with the consensual hypothesis. While there is no evidential basis to choose one plausible hypothesis over the other, the social support for the consensual hypothesis should make this hypothesis more plausible, and move participants to adopt it consequently. For determined rules, there is only one hypothesis compatible with the exemplars and the consensual hypothesis is implausible. Since the evidence is unambiguous, participants

who report the correct hypothesis may expect others to do so too. The disagreement with the other group members will disconfirm this expectation, which might result in conformity. However, no conformity was expected for the determined rules. The objective evidence was deemed enough ground for participants to reject the consensual position. Although in an anonymous Crutchfield-type experiment, in which only value-free judgements are made, the possibility for normative influence seems minimised, it cannot be ruled out beforehand. But, when no conformity is found for determined rules, it is plausible that conformity for the underdetermined rules is due to informational influence. Normative influences would affect the reported hypotheses regardless of the status of the ambiguity of the evidence.

4.1.1 Method

Participants and design

Seventy-one university undergraduates participated in the experiment, either in partial fulfillment of course requirements or for a small fee. There were 20 males and 51 females; the mean age was 22.91 ($SD = 6.57$).

The experiment had a 2 (no-consensus or consensus) $\times$ 2 (underdetermination or determination) factorial within-subjects design.

Procedure

Upon entering the experimentation room, participants were seated at a computer which was placed so that it was impossible to view the screens of the other participants. The computers were visibly linked through a network. Participants were informed the experiment concerned the differences between face-to-face and computer-mediated communication in a rule-discovery task and that they were placed in the latter condition. After this, they received extensive instructions concerning the inductive task. They were also told the two best performing participants would each receive a prize in the form of gift vouchers. Their performance would supposedly be assessed by the number of correctly chosen cards (the participants could keep track of this score) and the number of correct hypotheses given (participants could not keep track of this score). After a short practice trial consisting of a sequence of five exemplars, in which they received computerised feedback concerning the workings of the programme and their responses, they began with the four experimental trials. Each trial consisted of a different rule to discover, and was subject to one of the four experimental conditions. The order of these conditions was counterbalanced; there were four possible orders to which the participants were randomly assigned. After completion of the four trials, subjects answered eight exit-questions, inquiring whether they found the task difficult, how well they thought they performed in the task and whether they found the feedback of the others' hypotheses useful.

The inductive task involved the discovery of the rule that generated a sequence of playing-cards. This rule could be based on colour (red, black), suit (spades, hearts, clubs, diamonds), number (even, odd, 1, 2, ...), or a combination of number and colour or number and suit. Examples of such rules are 'all cards are even' or 'clubs followed by hearts followed by diamonds while the number of the next card is raised

by one'. After being shown the first three exemplars[1] of the rule, participants were asked to formulate a hypothesis concerning the rule by choosing from a set of possible hypotheses. They could either choose a hypothesis concerning colour or suit, one concerning number, or a combination of these two. There were 15 hypotheses related to colour or suit and 17 hypotheses related to number, resulting in a total of (15 + 17 + 15×17 =) 287 possible hypotheses. After reporting their hypothesis, for which they had thirty seconds, they were supposedly shown the hypotheses reported by the other three group members. Then they were shown a set of three playing cards, from which they had to choose the next exemplar of the rule. This choice was subject to a time limit of fifteen seconds. The time limits were set in order to synchronise the participants' answers so they would not get suspicious about the feedback. In a small pilot study, both time limits appeared long enough for participants to choose a hypothesis, while short enough for them not to get bored waiting. The card-choosing aspect of the task was included to control the amount of determination of the rule. By showing three cards, one of which was the next exemplar of the rule, the evidence concerning the rule could be precisely controlled. For instance, a card that was not the next exemplar of the rule could disconfirm a possible hypothesis held by the participant. If participants chose the correct card, they received a point. The point system was included as a motivational device. Then, regardless of whether they correctly identified the next exemplar, it was added to the existing sequence. Next, participants were asked to report their hypothesis concerning the rule based on this new information, after which they were shown the manipulated hypotheses of the other group members, etc. This cycle was repeated until the final, tenth exemplar of the rule was given, after which the participants were asked to give their final hypothesis. Thus, in each trial, participants were asked to report a total of eight hypotheses. The rules to be discovered were 'All cards have even numbers', 'The number of a card is first lowered by one, then raised by three', 'All cards are black' and 'Hearts follow spades, clubs follow hearts'.

Independent variables

Underdetermination

The cards from which the participants had to choose the next exemplar were chosen so as to result in either a determination or underdetermination of the rule. For a determined rule, the fourth or fifth exemplar in the sequence restricted the compatible hypotheses to one while for an underdetermined rule, multiple hypotheses were compatible with the sequence until the tenth exemplar. Due to the properties of a standard deck of cards, the multiple plausible hypotheses for an underdetermined rule can form a hierarchy. For instance, the sequences of exemplars that are compatible with 'hearts followed by diamonds' are a subset of those that are compatible with the rule 'all cards are red'. Clearly, when the rule 'hearts followed by diamonds' is the true generating rule, the rule 'all cards are red' will also be true. Participants were informed that in such a case, they should choose the hypothesis that corresponds with

[1]The term 'exemplar' shall be used throughout to denote particular cards that are instances of the rule. Each of the four trials consisted of the determination of the generating rule on the basis of a total of ten exemplars.

the true generating rule and not the one that is necessarily true when the other is true.

Consensus

The feedback about the others' hypotheses was manipulated so as to result in complete agreement (consensus) or disagreement (no consensus) among the three group members. In the no-consensus condition, group members reported different hypotheses after each consecutive exemplar, while in the consensus condition, all group members reported identical hypotheses after the fourth or fifth exemplar. In the no-consensus condition, there was always at least one person who reported the consensual hypothesis of the consensus condition, so that possible differences between these conditions would not be dependent on differences in information. The consensual hypothesis remained plausible until the last exemplar in the underdetermination condition. In the determination condition, the consensual hypothesis turned out to be inconsistent with the evidence after the fourth or fifth exemplar.

To clarify the nature of the task, an example of a trial for the different conditions is given in Appendix 4A.

Dependent variables

The analyses focus on the hypotheses reported by the participants. Due to the large number of possible hypotheses (287), it was necessary to categorise the reported hypotheses in three relevant classes.

Consensual Hypotheses (CH)

Each reported hypothesis was scored to indicate whether it was identical to the hypothesis constituting the consensus position for that rule. Since there was no consensual hypothesis in the no-consensus condition, hypotheses equal to the consensual hypothesis of the consensus condition do not indicate conformity. However, if the proportion of consensual hypotheses is larger in the consensus condition, as compared to the no-consensus condition, this indicates a conformity effect in the consensus condition.

Plausible Hypotheses (PH)

Each reported hypothesis was scored to indicate whether it was plausible, that is, consistent with the sequence of exemplars presented before the participant was asked to report the hypothesis.

Correct Hypotheses (CorH)

Each reported hypothesis was scored to indicate whether it was correct, that is, consistent with the entire sequence of exemplars.

Analysis

The experiment had a 2×2 factorial within-subjects design. Since, in the consensus condition, there was unanimity only in the last half of each trial, the reported hypotheses could be meaningfully divided into two blocks: those occurring before the formation of consensus and those occurring after the formation of consensus. For the analyses, we thus used an extra Block factor with two values (i.e. before and after), resulting in a 2×2×2 within-subjects design. For dependent variables with a normal

distribution, a 2×2×2 factorial repeated-measures ANOVA would be appropriate, but since the dependent variables were binary indicator variables, the distribution would clearly be non-normal. For binary variables, a logistic-regression model is more appropriate. This models the logit of the probability of a positive response as a linear function of the parameters. Since for a normally distributed variable, a regression analysis with dummy-coding for the experimental conditions is equivalent to an ANOVA analysis, the logistic-regression analysis with dummy-coding can be interpreted as the analogue of an ANOVA analysis for binary variables. To take into account the dependence of observations due to the repeated measurements of the same subjects, a random intercept was included for each subject. The random effects logistic regression analyses were performed using the MIXOR (Hedeker & Gibbons, 1996) programme, version 2.0. While it is plausible to assume a difference between the participants in their ability to report a plausible or correct hypothesis, and hence include a participant-specific random intercept in the analyses of PH and CorH, this does not make much sense for the analysis of consensual hypotheses. For each rule, the consensual hypothesis was a different hypothesis and there may have been a difference in the initial appeal of these hypotheses. Although participants may differ in their susceptibility to social influence, this should not result in a higher probability of reporting the consensual hypotheses over all four trials. Therefore, no random participant-specific intercept was included in the analysis for CH. In order to account for the possible difference in appeal between the consensual hypotheses, a fixed effect for the different rules was included.

4.1.2 Results

There were four possible orders of the conditions. There was no difference between the groups of participants who were assigned to one of these orders in the mean age, $F(3, 63) = 1.72$, $p = .17$, or the distribution of sex $\chi^2(3) = 1.12$, $p = .77$ [2].

Plausibility of the reported hypotheses

Since the social validation hypothesis assumes participants can distinguish between determined and underdetermined and with that plausible and non-plausible hypotheses, results concerning the plausibility of reported hypotheses are given first. The means and variances of PH are given in Table 4.1. No differences between the conditions were expected for the plausibility of the reported hypotheses, but respondents were expected to give a reasonable amount of plausible hypotheses. Overall, the proportion of PH was .69, indicating a reasonable level of difficulty in reporting hypotheses that are consistent with the evidence shown. As described in the method section, the Plausible Hypotheses were analysed with a logistic regression analysis with a participant-specific random intercept. The estimated standard deviation of the random intercept was $\hat{\sigma} = 1.39$ ($SE = .13, Z = 10.94, p < .001$), indicating a significant variability in individual performance. For the fixed part of the model, there were significant effects of Determination ($b = -1.83, SE = .21,\ Z = -8.82, p < .001$) and Block ($b = -1.15, SE = .27,\ Z = -4.23, p < .001$). Both parameter-estimates

[2] These analyses were based on $n = 63$, because the age and sex of eight participants were missing.

Table 4.1: Means, variances and predicted means of PH, CorH and CH[1]

	No consensus				Consensus			
	First		Last		First		Last	
	M	Var	*M*	Var	*M*	Var	*M*	Var
	Underdetermination							
PH	.88 (.93)	.11	.75 (.80)	.19	.84 (.93)	.13	.76 (.80)	.18
CorH	.04 (.02)	.03	.21 (.15)	.17	.06 (.02)	.06	.21 (.15)	.17
CH	.12 (.11)	.11	.17 (.14)	.14	.04 (.05)	.04	.23 (.25)	.18
	Determination							
PH	.64 (.67)	.23	.51 (.39)	.25	.63 (.67)	.23	.52 (.39)	.25
CorH	.31 (.24)	.22	.51 (.48)	.25	.34 (.24)	.22	.52 (.48)	.25
CH	.05 (.07)	.05	.01 (.04)	.01	.04 (.03)	.04	.09 (.06)	.08

[1] PH = Plausible Hypotheses, CorH = Correct Hypotheses, CH = Consensual Hypotheses. Values between parentheses are predicted proportions based on the fitted logistic regression models. Note that a binary variable with values 0 and 1, the mean equals the proportion of a positive value ($M = P(X = 1)$) and the variance equals the product of the proportions of the two values (Var $= P(X = 1)P(X = 0)$).

were negative, indicating that participants had a lower probability of reporting a plausible hypothesis when the rule was determined (as compared to underdetermined rules), and reported less plausible hypotheses in the last four responses than the first four. In logistic regression analysis with binary independent variables, the exponent of parameter-estimates corresponds to the estimated odds-ratio (Hosmer & Lemeshow, 2000). Looking at the parameter-estimates in this way, we see that participants were about 6.23 times more likely to report a plausible hypothesis for an underdetermined rule than a determined rule. Furthermore, participants were about 3.15 times more likely to report a plausible hypothesis in the first four responses than in the last four responses. However, as in normal regression analysis, caution must be applied when interpreting regression weights since they can be interdependent. To aid the interpretation of the effects of the independent variables on the probability of a plausible hypothesis, Table 4.1 contains the predicted probability of a plausible hypothesis in each cell of the design. These predictions were computed from a model which included only the intercept term and the significant effects of Determination and Block[3]. The main effect of Determination indicates that in the conditions in which the rule was determined, the reported hypotheses had a lower probability of being plausible. The main effect of Block indicates that the probability of a plausible hypothesis decreased from the first to the last four responses. Since the rules became more determined as more exemplars were added to the sequence, this reduction in probability from the first to last responses can also be attributed to a determination effect. Interestingly, these results indicate a relatively poorer performance as rules become more determined, which was not expected beforehand. A possible explanation is statistical: the

[3]Inclusion of only significant effects makes sense for two reasons. First of all, estimation of non-significant effects affects the estimation of the significant effects, leading to less reliable estimates of these latter effects. Second, incorporation of all effects leads to an exact replication of the observed mean probabilities, so that nothing is gained by investigating the the predicted probabilities.

probability that a randomly chosen hypothesis is plausible was higher in the low than in high determination condition. However, if subjects chose hypotheses on a random basis, then the number of correct hypotheses should not rise within the trials.

To ascertain whether the probability of a reported hypothesis being the correct one rose within trials, a logistic regression analysis with a participant-specific random intercept was applied to the Correct Hypotheses (CorH). A positive effect of Block (i.e. a rise in CorH from the first four to last four responses) was expected, as well as a larger probability of CorH in the determination than underdetermination condition. The estimated standard deviation of the random intercept was $\hat{\sigma} = 1.30$ ($Z = 6.13, p < .001$), indicating a significant individual variation in the ability to report the correct hypothesis. For the fixed part of the model, there were again significant effects of Determination ($b = 2.79, SE = .39,\ Z = 7.13, p < .001$) and Block ($b = 2.16, SE = .45,\ Z = 4.76, p < .001$). Participants were about 16.28 times more likely to report the correct hypothesis when the rule was determined by the evidence than when it was underdetermined. Participants were about 8.67 times more likely to report the correct hypothesis in the last four responses than in the first four responses. Finally, there was a significant Determination×Block interaction ($b = -1.08, SE = .52,\ Z = -2.07, p < .05$). For the interpretation of the combined effects, we shall again look at the predicted probabilities of a correct hypothesis as given in Table 4.1. These predictions were derived from a model incorporating only the significant Determination, Block and Determination×Block effects. The main effect of Determination indicates that the probability of reporting the correct hypothesis was higher in the determination condition than in the underdetermination condition. The main effect of Block indicates that in both conditions, the probability of a correct hypothesis rose within trials from the first to the last block. The interaction between Determination and Block finally indicates that this rise was highest in the determination condition. As for PH, all effects can be attributed to a determination effect. Concluding, we can say that the determination of the rules affected CorH as expected.

Conformity to consensual hypotheses

As indicated in the analysis-section, the Consensual Hypotheses (CH) were analysed with a logistic regression model with fixed effects for the conditions, as well as the four different rules. The results of this analysis are given in Table 4.2. As can be seen in Table 4.2, there were indeed significant differences between the rules in the probability of reporting the consensual hypothesis. Also, there were significant main effects of Consensus and Determination, as well a significant Consensus×Block and Determination×Block interaction. Both main effects were negative, indicating that participants were about 3.33 times more likely to report the consensual hypothesis in the no-consensus condition than in the consensus condition, and about 2.78 times more likely to report the consensual hypothesis in the underdetermination condition than in the determination condition. The negative influence of determination was expected, since the consensual hypothesis became implausible relatively early in the determination condition. The negative influence of consensus was not expected, but looking at Table 4.1, the effect seems attributable to initial differences (i.e. before

Table 4.2: Logistic regression for CH[1]

	b	$SE(b)$	Z	p	OR
(Intercept)	-1.06	0.22	-4.87	<.001	
Rule 2	-0.96	0.20	-4.75	<.001	0.38
Rule 3	-1.20	0.21	-5.62	<.001	0.30
Rule 4	-1.95	0.26	-7.42	<.001	0.14
Consensus (C)	-1.21	0.37	-3.27	.001	0.30
Determination (D)	-1.01	0.36	-2.81	.005	0.36
Block (B)	0.42	0.26	1.62	.106	1.52
C×D	1.02	0.59	1.74	.082	2.78
C×B	1.50	0.44	3.45	<.001	4.48
D×B	-2.04	0.71	-2.90	.004	0.13
C×D×B	1.03	0.89	1.16	.245	2.80

[1] OR = odds-ratio

the feedback showed consensus) between the no-consensus and consensus conditions. The Consensus×Block interaction shows that in the consensus condition, there was a rise in the probability of reporting the consensual hypothesis after the formation of consensus. The Determination×Block interaction shows a decrease in the probability of reporting the consensual hypothesis in the determination condition. Again, caution must be applied when directly interpreting regression weights. For better interpretation of the effects, the predicted probabilities for each cell of the design are given in Table 4.1. These predictions were derived from a model incorporating the significant Rule, Consensus, Determination, Consensus×Block and Determination×Block effects. The predicted probabilities confirm the earlier interpretation of the effects. The main effect of Consensus appears to be due to a higher initial probability of reporting the consensual hypothesis in the no-consensus condition. The main effect of Determination indicates a lower probability of reporting the consensual hypothesis in the determination condition than in the underdetermination condition. Furthermore, the predicted probabilities show that in all conditions apart from the no-consensus/determination condition, the probability of a consensual hypothesis rose within trials from the first to last block. The Consensus × Block interaction indicates that this rise was larger when there was consensus, while the Determination×Block interaction indicates that this rise was lower when the rule was determined. The slight rise in the predicted probability of a consensual hypothesis provides some indication of conformity to the consensual position in the consensus/determination condition. But, compared to the underdetermination condition, the level of conformity was very small.

4.1.3 Discussion

The results of experiment 1 were in agreement with the expectations from social comparison theory. Consensus resulted in conformity when the rule to be discovered was underdetermined by the evidence. This conformity was hardly found for determined

rules, and there seemed to be no general conformity effect. Therefore, the normative influence from the social feedback seems minimal and the conformity found for the underdetermined rules should be attributed to the informational influences. The results thus support Festinger's (1954) hypothesis that people rely on social comparison when objective evidence is ambiguous as to whether a belief is correct. Although the expectations from social comparison theory were supported, the overall level of conformity was rather moderate, with the highest predicted probability of a consensual hypothesis of .25 in the consensus/underdetermination condition. The effect of determination on the proposed hypotheses appeared much larger than that of consensus, so subjects seemed more influenced by the evidence than the hypotheses proposed by others. This is in agreement with Festinger's proposition that objective evidence is favoured for the evaluation of belief. The preference for information from the exemplars may also be due to particulars of the task. Although early on in the sequences, there was ambiguity as to which of the different plausible hypotheses is the correct one, by the final tenth exemplar the data always settled on a single hypothesis, so a reliance on social comparison would only have to be temporary. A stronger reliance on social validation might be expected when the level of underdetermination is higher, and underdetermination is pervasive rather than temporary.

4.2 Experiment 2

The goal of the second experiment was to study social comparison in situations of stronger underdetermination. When social comparison results from the ambiguity of the objective evidence, a stronger reliance on social comparison is expected as the belief becomes more underdetermined by the evidence. Subsequently, more conformity is expected for highly underdetermined rules than for rules with a lower level of underdetermination. The second experiment included confidence ratings, so the effects of social feedback on confidence could be investigated directly instead of through conformity behaviour. Although subjective confidence is considered an important variable in most theories on social influence, there have been few direct investigations of the influence of social support on confidence (McGarty, Turner, Oakes, & Haslam, 1993). Notable exceptions are the experiments of Goethals (1972); Goethals et al. (1978) and McGarty et al. (1993). These experiments showed that confidence is decreased by disagreement and increased by agreement and that, for beliefs, this increase is higher when agreement comes from a dissimilar rather than similar other. The present study is not so much concerned with the similarity thesis. Since purely factual beliefs are taken to be value-free, and all group members can base their hypotheses on the same evidence, participants will most likely assume similarity on attributes related to belief. However, if participants apply social projection (Orive, 1988), differences are expected for the different levels of underdetermination. In particular, the effect of social feedback on confidence is expected to differ for different levels of underdetermination. When a rule is less underdetermined, the judgement is less ambiguous and more agreement may be expected. As such, the informational value of agreement is lower when a rule is highly determined than when it is lowly determined. Thus, the effect of agreement on confidence is expected to be stronger the more underdetermined

a rule is, and agreement is expected to raise confidence more for highly underdetermined than for lowly underdetermined rules. For disagreement, the opposite is expected. When a rule is highly underdetermined, and hence the judgement more ambiguous, more disagreement may be expected. Therefore, the informational value of disagreement is lower the more underdetermined the rule becomes. When a rule is lowly underdetermined, less disagreement may be expected and the informational value of disagreement is higher. As such, disagreement should reduce confidence more when the rule has a lower level of underdetermination.

4.2.1 Method

Participants and design

79 university undergraduates participated in the experiment for partial fulfillment of course requirements or a small fee. There were 35 males and 44 females and the mean age was 22.30 ($SD = 3.50$).

The experiment had a 2 (consensus, no-consensus) $\times$ 2 (low underdetermination, high underdetermination) factorial within-subjects design[4].

Procedure

The procedure was similar to that of experiment 1. Participants entered the experimentation room in groups of four and were seated in front of computers which were connected through a network and placed so that it was impossible to see the screens of the other participants. The participants were first given instructions about the nature of the task and the manner in which they could report hypotheses. After receiving instructions concerning the inductive task, they started with the first of four trials.

The inductive task was slightly different from that of experiment 1. Due to the relation between suit and colour in standard playing cards, we found the manipulability of underdetermination too limited. Therefore, different cards were used in this experiment, consisting of three aspects: shape (square, triangle, circle, star), colour (red, blue, green or purple) and number (1 to 10). Again, the task consisted of reporting hypotheses for the rule that generated a sequence of these cards. To do this, subjects could choose one from a set of 33 hypotheses. The hypotheses referred to only one aspect of the cards, so either shape (for instance 'triangle follows circle'), colour (for instance 'all cards are red') or number (for instance 'the number of each consecutive card is raised by one'). Of the 33 offered hypotheses, 10 referred to shape, 13 to colour and 10 to number.

Each trial consisted of a generating rule to be discovered on the basis of a sequence of nine cards. Participants were first presented with three cards (the exemplars of the rule) and asked to report a hypothesis. They were then asked to rate their confidence in the reported hypothesis on a nine-point scale, ranging from 'completely unsure'

[4]The experiment also included a second manipulation of underdetermination, which was varied between subjects. In the condition with additional underdetermination, the instructions included a hint that some sequences might contain erroneous cards. In the condition without additional underdetermination, this hint was not included. Since this manipulation did not show any effect in any of the reported analyses, we will not discuss it further.

to 'completely sure'. After the first confidence rating, they received bogus feedback about the hypotheses of the other group members. Then the participants were asked to give a second confidence rating on the same nine-point scale. After adding two more exemplars to the sequence, participants reported a new hypothesis as well as their confidence in this hypothesis. Then they received social feedback and gave a second confidence rating. This was repeated twice more, each time after adding two more exemplars to the sequence. Thus, for each sequence of 9 exemplars participants were required to report a total of four hypotheses, one after the first three exemplars, one after the first five exemplars, one after the first seven and one after all nine exemplars.

Independent variables

Underdetermination

As in experiment 1, two levels of determination were included. In contrast to experiment 1, the rules remained underdetermined throughout the trials. In the low underdetermination condition, two hypotheses were consistent with the entire sequence. In the high underdetermination condition, three or four hypotheses were consistent with the entire sequence.

Consensus

As in experiment 1, the feedback about the others' hypotheses was manipulated so as to result in complete agreement (consensus) or disagreement (no consensus) among the three group members. In the no-consensus condition, group members reported different hypotheses after each consecutive exemplar, while in the consensus condition, all group members reported identical hypotheses after the fifth exemplar. In the no-consensus condition, there was always at least one person who reported the hypothesis that was unanimously endorsed in the consensus condition.

Dependent variables

Since all rules remained underdetermined throughout the trials, the designation of one plausible hypothesis as 'correct' is arbitrary. Therefore, the CorH variable was not used in this experiment. In addition to the Plausible Hypotheses (PH) and Consensual Hypotheses (CH) variables as used in experiment 1, there were two confidence variables, one for the confidence ratings before receiving the social feedback (Conf1) and one for the ratings after (Conf2).

4.2.2 Results

Plausibility of the reported hypotheses

As before, the results for PH are discussed first. The means and variances of the plausible hypotheses (PH) for the different conditions can be found in Table 4.3. Overall, the proportion of PH was .94, indicating that the task was relatively easy as compared to the first experiment, in which the overall proportion of PH was .69. As in experiment 1, the plausible hypotheses were analysed with a logistic regression model with a participant-specific random intercept. The estimated standard deviation of

Table 4.3: Means and variances of PH and CH[1]

	No consensus				Consensus			
	First		Last		First		Last	
	M	Var	*M*	Var	*M*	Var	*M*	Var
	High underdetermination							
PH	.95 (.93)	.05	1.00 (.99)	.00	.95 (.93)	.05	.98 (.99)	.02
CH	.25 (.29)	.19	.33 (.29)	.22	.29 (.38)	.21	.47 (.38)	.25
	Low underdetermination							
PH	.92 (.93)	.07	.86 (.95)	.12	.88 (.93)	.11	.93 (.95)	.07
CH	.44 (.56)	.25	.67 (.56)	.22	.35 (.49)	.23	.61 (.49)	.24

[1] PH = plausible hypothesis, CH = consensual hypothesis. Values between parentheses are predicted proportions based on the fitted logistic regression models.

the random intercept was $\hat{\sigma} = 1.24$ ($SE = .18, Z = 6.79, p < .001$), indicating a significant variability in participants' ability to report a plausible hypothesis. The fixed part of the model showed a significant effect for Block ($b = 2.03, SE = .71$, $Z = 2.86, p < .01$) and a significant Underdetermination×Block interaction ($b = -1.76, SE = .89$, $Z = -1.98, p < .05$). The positive parameter estimate of Block indicates a rise in the probability of a plausible hypothesis from the first two to the last two responses. Participants were about 7.61 times as likely to report a plausible hypothesis in the last two responses than the first two. The parameter estimate of the Underdetermination×Block effect was negative, indicating that this rise was less for lowly underdetermined than highly underdetermined rules. To see how these two effects affected the predicted probabilities of PH for each of the cells in the design, Table 4.3 contains the predicted probability of PH, based on a model with only the significant Block and Underdetermination×Block effects. As can be seen, the predicted probabilities were relatively high in each cell. While the probability of a plausible hypothesis rose from the first to second block, indicating that as a rule became more determined, it was easier to report a plausible hypothesis, the main effect for Underdetermination was in the opposite direction, showing a higher probability of a plausible hypothesis for the highly underdetermined rules. Again, a simple explanation is found in the fact that there were more plausible hypotheses for the highly underdetermined rules.

Conformity to consensual hypotheses

The means and standard deviations of Consensual Hypotheses (CH) are given in Table 4.3. As in experiment 1, a rise in the probability of a consensual hypothesis from the first to the last two responses indicates conformity to the consensual position in the consensus condition. In other words, conformity to the consensual hypothesis is indicated by a positive Consensus×Block interaction. As in experiment 1, the consensual hypotheses were analysed using a logistic regression model with fixed effects for Consensus, Underdetermination, Block and their interactions, as well as fixed main effects for the different rules. This analysis showed only a significant

effect for Underdetermination ($b = .85, SE = .26,\ Z = 3.32, p < .001$), indicating that participants were about 2.34 times as likely to report the consensual hypothesis when the rule was lowly underdetermined. This effect was expected, since in this condition the consensual hypothesis was one of two plausible hypotheses, while in the high underdetermination condition it was one of three or four plausible hypotheses. There were no significant effects for the different rules, indicating that the different consensual hypotheses did not differ much in their initial appeal. There was also no significant Consensus×Block interaction and hence no indication of a conformity effect. Since there were only two responses in each block, as compared to four in experiment 1, there may have been a loss in statistical power. Although it does not provide direct evidence for conformity, a logistic regression model without the Block effect did result in a significant effect for Consensus ($b = 0.40,\ SE = .17,\ Z = 2.30,\ p = .022$) and Underdetermination ($b = 1.13,\ SE = .18,\ Z = 6.42,\ p < .001$), as well as a significant Consensus×Underdetermination interaction ($b = -0.71,\ SE = .24,\ Z = -2.94,\ p = .003$). The effects were in the expected directions. Participants were about 1.49 times more likely to report the consensual hypothesis when it was supported by a unanimous group than when only one other group member supported this hypothesis. Due to the exclusion of the Block effect, the consensus effect cannot be immediately interpreted as a conformity effect, because there may have been more participants initially (before receiving social feedback) reporting the consensual hypothesis in the consensus condition. However, this was not the case. The proportion of participants initially reporting the consensual hypothesis in the consensus condition was .19, which is actually lower than the no-consensus condition, where this proportion was .26, although the difference was not significant $\chi^2(1) = 1.93,\ p = .17$. Since the consensus effect cannot be attributed to initial differences, it can be attributed to a rise in the number of participants reporting the consensual hypothesis after they received social feedback, thereby indicating a conformity effect. As in the first experiment, the effect of underdetermination seemed stronger than that of consensus: participants were about 3.10 times more likely to report the consensual hypothesis when it was one of two plausible hypotheses rather than one of three or four plausible hypotheses. The Consensus×Underdetermination effect indicates that the effect of consensus was less when the rule was less underdetermined by the evidence. The predicted probabilities derived from the logistic regression model with Consensus, Underdetermination and Consensus×Underdetermination effects are given in Table 4.3. These predicted probabilities show a puzzling effect of the Consensus×Underdetermination interaction. For highly underdetermined rules, the probability of a consensual hypothesis was higher in the consensus condition than in the no-consensus condition. For rules with a low level of underdetermination, this relation was reversed, so there was a higher probability of reporting the consensual hypothesis in the no-consensus condition than the consensus condition. This could indicate a resistance to informational influence from the group when belief was only mildly underdetermined by the evidence. However, it may have also been due to a difference in the number of participants initially reporting the consensual hypothesis. In the low-underdetermination/no-consensus condition, the proportion of participants initially reporting the consensual hypothesis was .28, while this proportion was only .18 in the low-underdetermination/consensus condition. Although this difference was

Table 4.4: Means and variances of Conf1 and Conf2[1]

	No consensus				Consensus			
	First		Last		First		Last	
	M	Var	*M*	Var	*M*	Var	*M*	Var
	High underdetermination							
Conf1	6.15	4.81	7.14	3.45	6.05	4.55	7.34	3.15
Conf2	5.97	4.61	7.00	4.16	5.95	5.16	7.40	3.10
	Low underdetermination							
Conf1	5.99	5.81	6.97	4.47	5.94	4.82	7.27	4.15
Conf2	5.98	6.26	7.03	4.79	5.86	5.85	7.34	4.29

[1] Conf1 = confidence rating before social feedback, Conf2 = confidence rating after social feedback.

again not significant, $\chi^2(1) = 1.56$, $p = .21$, it may have affected the interaction effect. The overall proportion of participants reporting the consensual hypothesis in the low-underdetermination/no-consensus condition (.56) was not significantly higher than the corresponding proportion in the low-underdetermination/consensus condition (.49), $\chi^2(1) = 3.30$, $p = .07$. In the high-underdetermination condition, there was a significant difference for this proportion between the consensus and no-consensus condition, $\chi^2(1) = 5.38$, $p = .02$. Although the interaction may have been due to a resistance to informational influence in the low-underdetermination condition, the evidence for such resistance is inconclusive. While the possible resistance effect needs further investigation, the hypothesis that conformity is higher for highly underdetermined than lowly underdetermined rules was clearly supported.

Confidence

For each reported hypothesis, two confidence ratings were given: one before, and one after receiving social feedback regarding the hypotheses of the other group members. We shall refer to these as Conf1 and Conf2 respectively. The means and variances of the two confidence ratings are given in Table 4.4. No direct effects of the experimental manipulations were expected on the overall level of confidence. Confidence was expected to be more related to the hypotheses held, and effects of the experimental manipulations were expected only in interaction with hypothesis type. To detect possible overall differences, which may have influenced the results of later analyses, the confidence ratings were first analysed with an univariate repeated-measures ANOVA, treating the two confidence scores as repeated measures of the same variable, distinguished by a dummy variable to indicate whether the confidence was given before or after the social feedback. This 2×2×2×2 repeated-measures ANOVA resulted in a significant effect only for Block, $F(1, 66) = 154.93, p < .001$, indicating that the confidence in the first two reported hypotheses ($M = 5.99$, $SD = 2.28$) was lower than the confidence in the last two reported hypotheses ($M = 7.19$, $SD = 1.99$). This analysis does not distinguish between hypothesis types (i.e. plausible or consensual). Overall, we expected the confidence in a plausible hypothesis to be higher than the confidence

in an implausible hypothesis. Since there was a significant difference between the means of Conf1 for a non-plausible hypothesis ($M = 4.92, SD = 2.72$) and a plausible hypothesis ($M = 6.84, SD = 1.97$), $t(81) = -6.07, p < .001$, this appeared to be the case.

Following Goethals et al. (1978), the effect of social feedback on the confidence in a hypothesis was investigated by means of an ANCOVA for the second confidence ratings (Conf2), with the first confidence ratings (Conf1) as a covariate. The included fixed effects were Consensus and Underdetermination, as well as CH, the binary variable indicating whether the hypothesis was the consensual hypothesis or not. We expected a significant interaction between Consensus and CH, since in the consensus condition, there was unanimous support for the consensual hypothesis. This support should have lowered confidence in a non-consensual hypothesis, and raised the confidence in a consensual hypothesis. To test this hypothesis, the confidence ratings for non-plausible hypotheses were excluded from the analysis (due to the large percentage of plausible hypotheses, only 6% of the cases needed to be excluded). The reason for this is that the effect of CH can be inflated when the non-consensual hypotheses can be either plausible or non-plausible (the consensual hypothesis was always plausible), since plausibility was related to confidence. As expected, the ANCOVA showed a significant Consensus×CH interaction, $F(1, 1051) = 17.40$, $p < .001$. Looking at the adjusted means (the means of Conf2 as deviations from Conf1), we see that in the consensus condition, there was a relatively large difference between the change in confidence in the consensual hypothesis ($M = .30, SD = .06$) and in a non-consensual but plausible hypothesis ($M = -.26, SD = .06$). In the no-consensus condition, there was hardly a difference between the change in confidence in the consensual hypothesis ($M = -.01, SD = .06$) and a non-consensual but plausible hypothesis ($M = -.03, SD = .06$). Besides the expected interaction effect, there was also a main effect of CH, $F(1, 1051) = 20.54$, $p < .001$, indicating an overall increase in confidence in the consensual hypothesis ($M = .15, SD = .04$) and an overall decrease in the confidence in a non-consensual but plausible hypothesis ($M = -.14, SD = .04$). Finally, there was a significant Consensus×Underdetermination interaction, $F(1, 1051) = 7.36$, $p = .007$. This effect indicates that while confidence increased in both the high-underdetermination/consensus ($M = .07, SD = .06$) and low-underdetermination/no-consensus ($M = .10, SD = .06$) conditions, there was an overall decrease in the high-underdetermination/no-consensus ($M = -.14, SD = .06$) and low-underdetermination/consensus ($M = -.02, SD = .06$) conditions. This interaction is difficult to interpret, since it does not incorporate the distinction between hypothesis types.

As indicated earlier, a particular interaction between agreement and the level of underdetermination was expected. In the above analysis, no significant Consensus × Underdetermination × CH interaction was found. A more direct test of the expectation was obtained by applying a contrast to the ANCOVA. Changes in confidence following social feedback were only expected in the consensus condition. So, in the absence of consensus, no effects of Determination or CH on Conf2 were expected. In the consensus condition, we expected that disagreement (i.e. the participant reporting a non-consensual hypothesis) would result in a decrease in confidence, which is larger for the low-underdetermination condition than for the high-underdetermination con-

Table 4.5: Adjusted means of Conf2[1]

Underdetermination:	High		Low	
	M	SD	M	SD
		No consensus		
Non-consensual hypothesis	-0.10	0.07	0.05	0.10
Consensual hypothesis	-0.16	0.10	0.15	0.08
		Consensus		
Non-consensual hypothesis	-0.17	0.07	-0.34	0.09
Consensual hypothesis	0.30	0.09	0.30	0.08

dition. Agreement was expected to result in an increase in confidence, which should be higher in the high-underdetermination condition than in the low-underdetermination condition. To test this hypothesis, a contrast was specified with coefficients equal to 0 for the no-consensus conditions. For the consensus conditions, coefficient values of -1, 2, -2, and 1, were used for the high-underdetermination/disagreement, high-underdetermination/agreement, low-underdetermination/disagreement and low underdetermination/agreement combinations respectively. This contrast described differences between the adjusted means of Conf2 very well, $F(1, 1051) = 42.18$, $p < .001$, and the residual variance was non-significant, $F(7, 1051) = 1.88$, $p = .071$. The expected pattern was clearly confirmed, but planned comparisons did not show a significant difference between the high-underdetermination/disagreement and low-underdetermination/disagreement cells, nor between the high-underdetermination/ agreement and low-underdetermination/agreement cells. From the adjusted means of Conf2, given in Table 4.5, it appears that although disagreement (non-consensual hypothesis) resulted in a larger decrease in confidence for highly than lowly underdetermined rules, there was no difference between the levels of underdetermination in the rise in confidence following agreement (consensual hypothesis). Thus, although the contrast fitted the results very well, the expectation that agreement would result in a larger increase in confidence when the hypothesis was highly underdetermined was less supported than the expectation that disagreement would result in a larger decrease in confidence when the hypothesis was lowly underdetermined.

4.2.3 Discussion

Experiment 2 did not show direct evidence for conformity to the consensual position. Indirect evidence was found however, since overall the consensual hypothesis was reported more in the consensus condition than in the no-consensus condition, a result which could not be attributed to a higher proportion of participants initially reporting this hypothesis in the consensus condition. Also, as expected, there was more evidence for conformity in the low determination condition than in the high determination condition. The effect of consensus showed itself more clearly in the analysis in participants' confidence in their reported hypotheses. The unanimous support for the consensual hypothesis raised the confidence in this hypothesis, and lowered the confi-

dence in a non-consensual hypothesis. Moreover, the decrease in confidence following disagreement appeared higher when the rule was less underdetermined. Although disagreeing with a unanimous group did lower confidence in a hypothesis, especially in the low underdetermination condition, this may not have been enough to persuade participants to abandon their own hypothesis in favour of the consensual one.

4.3 General discussion

The two experiments presented support Festinger's basic claim that people evaluate beliefs by social comparison when the objective evidence is insufficient for this purpose. In the first experiment, strong evidence was found for the informational influence of an unanimous group belief on the belief of individual participants. In a situation of underdetermination, where there is equal objective support for a consensual hypothesis and another plausible hypothesis, the social support for the consensual hypothesis resulted in a conformity to this position. In the second experiment, the level of underdetermination was varied in order to test the hypothesis that the informational influence of the group is stronger when the objective evidence is more ambiguous. Although there was less direct evidence for conformity in this experiment, this hypothesis was supported. Stronger evidence for the social comparison of belief came from the effects of social feedback on the confidence in reported hypotheses. Confidence was lowered when participants reported a hypothesis that differed from one unanimously endorsed by the other group members, while confidence was raised when participants reported a hypothesis that was identical to the consensual hypothesis. Furthermore, the effects of agreement or disagreement on confidence appeared to differ for the different levels of underdetermination. It was expected that the increase in confidence following agreement would be stronger for highly than for lowly underdetermined rules, while the decrease in confidence following disagreement would be stronger for lowly underdetermined than for highly underdetermined rules. This expectation was supported, although most clearly for the decrease in confidence following disagreement.

That the first experiment provided stronger evidence for conformity than the second experiment may be due to the relative easiness of the task in the latter experiment. Participants in the first experiment had more difficulty reporting a plausible hypothesis than in those the second experiment. This difficulty may have resulted in a lower overall confidence and hence more impetus for social comparison to reduce this uncertainty. If participants were more confident in the second experiment, disagreeing with an unanimous group may not have lowered confidence enough to result in a change of belief. Interestingly, while the decrease in confidence following disagreement was higher in the low underdetermination condition, there was less evidence for conformity in this condition as compared to the high underdetermination condition. While participants may have assumed more agreement in the low underdetermination condition, and the disconfirmation of this expectation resulted in a relatively large decrease in confidence, participants were less inclined to resolve this uncertainty by adopting the consensual hypothesis. A possible explanation for this is that in the low underdetermination condition, there were two hypotheses consistent with the evidence, while

in the high underdetermination condition, there were three or four. The probability that a plausible hypothesis is correct is therefore higher in the low underdetermination than in the high underdetermination condition. For this reason, dissenters in the low underdetermination may have been more certain about the correctness of their plausible hypothesis than dissenters in the high underdetermination condition.

Another difference between the two experiments is that in the first experiment the entire sequence of exemplars was always conclusive evidence in favour of a single generating rule. In the second experiment the rules remained underdetermined. This higher level of ambiguity in the second experiment was expected to result in a higher reliance of social comparison and, as a result, a higher level of conformity. On the other hand, the ambiguity may have interfered with the participants' perception of the task as principally objective. Although the task was considered objective in both experiments, being concerned with potentially verifiable statements concerning the generating rule behind a sequence of data, only in the first experiment was this potential of verifiability realised. The pervasive underdetermination in the second experiment may have led some participants to believe there was no objectively correct hypothesis, or at least not one they could ever determine. Consequently, these participants would not be expected to use social comparison, for why attempt to evaluate the correctness of a belief if there is no correct belief? The assumption of objectivity, i.e. the possibility of a true belief, seems critical for the social comparison of belief. Evidence for this was found in an experiment by Insko et al. (1983). They studied conformity in a task in which participants had to indicate which of two colours was more alike a third colour. In one group, participants were led to believe that the correct answer could be objectively determined by means of a spectrometer. In the other condition, participants were led to believe that this was impossible. As the authors expected, conformity was much stronger in the first condition than in the latter. According to Insko et al., participants in the first condition were more concerned with being right than in the second condition, and hence more persuaded to apply social comparison. The results are consistent with the idea that the informational value of others' beliefs is higher when the task is objective rather than subjective. Similar results were obtained by Olson et al. (1983). Apparently, the occurrence of social comparison hinges on two factors: (1) the assumption of a single true belief, and (2) the underdetermination of this belief by objective evidence. Both aspects are related to the informational value of others' beliefs. When there is a single correct position, all beliefs pertain to this single object, so that any belief has informational value in principle. When the correct position is underdetermined by objective evidence, there are multiple plausible positions, and there is no objective ground to favour one over the other. In order to make a trustworthy choice, the social support may then be used as additional evidence to discriminate among the otherwise equally plausible positions.

While social comparison did appear to influence the individual hypotheses in both experiments reported here, the effect of objective information on the formation and change of belief appeared much larger. This is in agreement with Festinger's assumption that people prefer objective evidence for the evaluation of belief. In inductive tasks, where there is a single true rule that generated the data, an increase in the amount of evidence has been shown to raise the number of correct hypotheses more

than an increase in the number of proposed hypotheses (Laughlin, 1999; Laughlin & Bonner, 1999). The correct hypothesis must be one of all possible hypotheses that are consistent with the evidence. Enlarging the set of explicitly proposed plausible hypotheses will raise the probability that the correct hypothesis is a member of this set. However, enlarging the set of plausible hypotheses will not make the correct hypothesis stand out more, nor raise the probability that a (randomly) chosen hypothesis from the set of proposed plausible hypotheses is correct. Only an increase in the amount of evidence, and with that a decrease in the amount of possible plausible hypotheses, can accomplish this.

Social comparison can be a valuable means of reducing effects of idiosyncratic biases in availability, selection, and processing of information on belief formation. In the two experiments reported here, all participants based their hypotheses on the same information. For this reason, disagreement should be attributed to differences in the processing of this information, not in differences in the amount of available information. Suls et al. (2000) found a preference for dissimilar advantaged others (i.e. experts or persons who possess more information) in the comparison of belief, supporting the idea that people use social comparison in order to counter bias stemming from insufficient information. The results of the two experiments described here indicate an additional motivation to counter idiosyncratic bias in the processing of information.

Appendix

4A Example of a trial in experiment 1

This appendix contains an example of a trial in experiment 1. A sequence of 10 exemplars is given for both the Underdetermination and Determination conditions. Immediately below the sequence are the two cards which, together with the exemplar, were the set of three cards from which participants had to choose the next exemplar. Below the objective evidence is an overview of the plausible hypotheses, i.e. the hypothetical rules that are consistent with the sequence of exemplars given so far. Note that these lists start at exemplar 4, since this is where participants were to give their first hypothesis. Hypotheses are abbreviated, with 'even' denoting the hypothesis 'All cards have an even number', 'R B B' denoting the hypothesis 'A Red card is followed by two Black cards' (which are followed by another red card and two black cards, etc.), and '+0 -2' denoting the hypothesis 'A card is followed by a card with the same number, and then followed by a card with a number lowered by two' (which is then followed by a card with the same number and one with a number -2, etc). Below the sequences and plausible hypotheses, the feedback is given for the no-consensus and consensus conditions.

	1	2	3	4	5	6	7	8	9	10
					Underdetermination					
Exemplar:	♡10	♠10	♣8	♡8	♣6	♠6	♡4	♢4	♠2	♠8
Other cards:				♣7	♢7	♢5	♠3	♠1	♢1	♡3
				♠9	♡5	♡7	♠9	♠5	♠3	♠1
Plausible hypotheses				even +0 −2 R B B ♡♠♣	even +0 −2 R B B	even +0 −2	even +0 −2	even +0 −2	even +0 −2	even
					Determination					
Exemplar:	♡10	♠10	♣8	♢8	♡2	♣6	♡4	♠4	♣2	♡8
Other cards:				♣7	♢3	♠5	♢9	♠1	♢7	♡1
				♠9	♣1	♡7	♣3	♡5	♠3	♣9
Plausible hypotheses				even +0 −2	even	even	even	even	even	even
Feedback consensus				♡♠♣ R B B +0 −2	♡♠♣ R B B +0 −2	R B B +0 −2 +0 −2	+0 −2 +0 −2 +0 −2	+0 −2 +0 −2 +0 -2	+0 −2 +0 −2 +0 -2	
Feedback no-consensus				♡♠♣ +0 −2 R B B	♡♠♣ +0 −2 R B B	♡♠♣ +0 −2 +0 −2	♡♠♣ even +0 −2	♡♠♣ even +0 −2	♡♠♣ even +0 −2	

5
Collaboration in nonmetric multiple cue probability learning

There are at least two reasons why a group can outperform individuals when it comes to making good judgements and decisions (Meehl, 1999). The first is that individuals may possess (partly) non-overlapping information, so that the group as a whole can base its judgement or decision on more information than any individual alone. The second is that idiosyncratic biases may affect the group judgement or decision to a lesser extent than individual ones, because the idiosyncratic biases may cancel each other out, so to say, in a group judgement or decision. While these reasons render it plausible that groups are advantaged over individuals, previous research has shown that groups often do not realise their potential. For instance, it has been shown that group discussions focus mainly on information shared by group members, rather than on the unique information individuals can contribute (Stasser & Titus, 1985; Gigone & Hastie, 1993; Larson, Christensen, Abbott, & Franz, 1996). Even when unique information is discussed, the influence of this information on the final group judgement or decision is much smaller than that of shared information. The preference for shared over unique information has been labelled the 'common knowledge effect' (Gigone & Hastie, 1993), and it shows that while groups are potentially advantaged over individuals, they usually do not realise this potential fully. As for the 'cancelling-out' of idiosyncratic bias argument, research shows that some biases which are often found on an individual level, such as the base-rate fallacy, may actually be exaggerated in a group decision (Tindale, 1993; Kerr, MacCoun, & Kramer, 1996; Kerr, Niedermeier, & Kaplan, 1999). Findings such as these show that the assumption that groups arrive at better judgements and decisions should be scrutinised. The purpose of this chapter is to do so in the context of group decision-making based on probabilistic information, an area in which group performance has not been studied extensively.

Overall, it has been an ubiquitous finding that actual group performance lies below potential group performance, i.e. the performance that would be obtained if a group makes optimal use of its members' resources (Kerr & Tindale, 2004). In Steiner's (1972) terminology, groups exhibit process loss. Steiner takes group performance (he uses the term group productivity) to depend on three components: task demands, group resources and group process. Task demands specify the resources needed, and how they should be combined, in order to complete the task. Group resources are those resources available to the group as a whole. Together, these two components determine the potential performance of the group. The group process refers to the process through which the group combines the group resources and determines to what extent the potential performance is realised. The actual group performance depends on the match between the task demands on the one hand, and the group resources and process on the other. Task demands, which are not under the control of the group, are critical determinants of both potential and actual group performance. They form the basis of Steiner's (1972) typology of group tasks. Tasks are distinguished on three dimensions. The first dimension concerns whether the task is divisible or unitary. Divisible tasks can be divided into subtasks in such a way that a group can achieve maximal performance if every member completes just one subtask. For unitary tasks, a profitable division into subtasks is not possible. The second dimension concerns whether the task is of a maximising or optimising kind. Maximising tasks require as much as is possible of something, such as force in a rope-pulling contest or ideas in a brainstorming session. In optimising tasks, the goal is to produce a specific, correct or most preferred outcome. Decisions and judgements are optimising tasks. The third dimension concerns the permitted and prescribed group process. This dimension distinguishes between four task-types: additive, disjunctive, conjunctive and discretionary tasks.

In *additive* tasks, group performance equals the sum of the individual contributions. The rope-pulling contest is an example of an additive task, since the force a group exerts is the sum of the force of each group member. While the group performance in an additive task must always be at least as high as the actual performance of any individual member, this is not to say that group performance is necessarily higher than the potential performance of any of its members. Classic research by Ringelman (cited in Steiner, 1972) showed that the force a group exerts is smaller than the sum of the forces the group members exert when pulling individually. This can be due to a problem in coordination, for instance if not all group members pull at the same time, or a problem in motivation, if group members exert less force because they are working in a group. In disjunctive and conjunctive tasks, the group product must equal the contribution of a single group member. In *disjunctive* tasks, the group is free to choose which contribution to adopt as its own. Group performance is optimal if the group adopts the contribution of the most proficient member. In unitary disjunctive tasks, groups can only perform successfully if at least one member possesses all resources necessary to complete the task. An important class of disjunctive tasks are so-called Eureka-problems (Lorge & Solomon, 1955). These are problems in which the correctness of the proper solution, once obtained, can be easily demonstrated. Demonstrability is a key characteristic of intellective tasks (Laughlin & Ellis, 1986). For such tasks, it has been shown that groups mostly perform at the level of the best

group member. The advantage of the group lies in the simple fact that as the size of the group increases, so does the probability that at least one member possesses the necessary resources. When the demonstrability requirement is not met, groups often do not perform at the level of the best individual member. In *conjunctive* tasks, the group is not free to choose which member's contribution to adopt as the group product. Performance in these tasks is necessarily identical to the performance of the least proficient member. A climbing expedition is an example of such a task, since the expedition can only move at the speed of its slowest member. *Discretionary* tasks comprise the last task type. In these tasks a group is free to combine the individual contributions in any way it sees fit. Judgement and decision tasks are usually of this type.

Most judgements, and with that the decisions which follow from these judgements, are based on information which is not deterministically related to the subject of judgement. Furthermore, this information may be the only indication of the true state of the object to which the judgement pertains. In Brunswik's (1955) terms, it is a situation in which a distal criterion can only be perceived through proximal cues, which are probabilistically related to the criterion. Brunswik's theory of probabilistic functionalism addresses how an organism adapts to the probabilistic relations in its environment. This process is directly investigated in Multiple Cue Probability Learning (MCPL) tasks, in which individuals learn to predict the value of a criterion on the basis of a number of cues. Brunswik's lens model provides the framework for analysing performance in such tasks. The lens model concerns how information from multiple cues is combined into individual judgements, and how the relation between cues and judgements compares to the relation between cues and an objective criterion. In the so-called two system approach, the proximal cues and distal criterion comprise the *ecological system* and the cues and judgements the *judgement system*. Hammond, Wilkins, and Todd (1966) proposed to extend the two system approach to a three system approach, with one ecological system and two judgement systems, or even to an n-system approach. These extensions provide a framework in which social processes in judgement can be precisely studied. Such research goes under the name of Interpersonal Conflict (IPC) or Interpersonal Learning (IPL). The basic structure of both is as follows. Two individuals participate in a MCPL experiment with an ecological system consisting of two cues. The first individual is trained in a system where cue 1 is strongly related to the criterion, but cue 2 is not related. The second individual is trained in a system where cue 1 is not related to the criterion, while cue 2 is strongly related. In the interpersonal learning stage, the individuals are put together to learn in a system where both cues are related to the criterion, although less than before. By learning from each other, the participants are expected to perform better in this new system than by working alone. Research conducted using this paradigm (Brehmer, 1973, 1974) has shown that individuals often tend to change an optimal judgement procedure toward a less optimal one under interpersonal learning conditions. Problematic in this approach is that the ecological system changes between the individual learning stage and the interpersonal learning stage. This will probably be quite confusing for the participants, and such sudden changes in the ecological system are not expected in realistic situations where interpersonal learning might take place. In fact, the IPL studies as previously conducted required

group members to unlearn part of what they previously learned, rather than learning about aspects of the ecological system from each other. Such unlearning will be much more difficult than learning about aspects of the ecological system from each other's responses. Indeed, research by Andersson and Brehmer (1979) has shown that the diminished individual performance in IPL conditions is attributable to the change in the ecological system, rather than to the social environment.

A way to overcome this problem of IPL research is to train participants in different parts of the same ecological system. In this way, each group member has knowledge of only a partial ecological system, but, depending on how this knowledge is distributed in the group, the group as a whole may have more knowledge about the complete ecological system than any group member alone. If each group member bases his response on a different partial ecological system, the group can maximise the information it possesses as a whole. Such a division of cognitive labour is effective if there is a limit to the amount of information that individuals can process adequately and individual contributions are properly integrated in the group process. The two experiments described here investigate such collaboration in Nonmetric Multiple Cue Probability Learning (NMCPL) tasks under different distributions of information over group members. In the following, a brief overview of NMCPL is given, as well as methods for the analysis of individual behaviour in such tasks. Then, a group paradigm is described, as well as possible group processes in such collective NMCPL tasks.

5.1 NMCPL

In nonmetric multiple cue probability learning (NMCPL) tasks, the objective is to learn about the probabilistic relations between a categorical event e_k, $k = 1, \ldots, K$ and a number of categorical cues $c_j, j = 1, \ldots, J$, in order to predict the event on the basis of observed cues. A common NMCPL task is that of medical diagnosis, where a physician has to decide whether a disease is present on the basis of the presence of a number of symptoms. NMCPL tasks are a special case of general multiple cue probability learning tasks and can be analysed in a similar vain as metric MCPL tasks, using the lens model approach. For metric MCPL tasks, achievement is defined as the correlation between judgements r_i and event e_k. The lens model equation (Tucker, 1964) shows how this correlation is a function of four components, which have been termed the predictability of the event, the cognitive control of the individual, the individual's linear knowledge and his configural knowledge (e.g. Cooksey, 1996). For nonmetric MCPL tasks, achievement is more naturally defined as $P(r_i = e_k)$, the probability that the response is identical to the event. A nonmetric version of the lens model equation has been derived by Björkman (1973) for an ecological system consisting of a polytomous criterium and a single polytomous cue. It is easy to generalise his approach to an ecology with more than one cue, by analysing the multiple cue environment as a single cue environment, in which the single cue C represents the cartesian product of the individual cues c_j. For example, for an environment with two dichotomous cues, the variable C would have $2^2 = 4$ levels: $C_1 = (c_{1:1}, c_{2:1})$, $C_2 = (c_{1:1}, c_{2:0})$, $C_3 = (c_{1:0}, c_{2:1})$, and $C_4 = (c_{1:0}, c_{2:0})$.

The main elements of a double system design for NMCPL are the cue profiles C_m,

the event e_k and the responses r_i. From the standard axioms of probability, the joint probability $P(r_i, e_k, C_m)$ can be partitioned as $P(r_i, e_k|C_m)P(C_m)$. But since we can assume the responses and event are conditionally independent, we have:

$$P(r_i, e_k, C_m) = P(r_i|C_m)P(e_k|C_m)P(C_m). \tag{5.1}$$

5.1.1 Achievement

The achievement A_i of individual i is defined as

$$A_i \equiv P(r_i = e_k) = \sum_k \sum_m P(r_{i:k}|C_m)P(e_k|C_m)P(C_m). \tag{5.2}$$

An analogous formulation can be given in matrix algebra. While the approach is general, it will be illustrated here for a simple system with a dichotomous criterium e, with values e_1 and e_{-1}, and two dichotomous cues c_j, with values $c_{j:1}$ and $c_{j:0}$. Let $\boldsymbol{c}$ be a vector of the base-rate probabilities of the cue profiles

$$\boldsymbol{c} = \begin{bmatrix} P(c_{1:1}, c_{2:1}) \\ P(c_{1:1}, c_{2:0}) \\ P(c_{1:0}, c_{2:1}) \\ P(c_{1:0}, c_{2:0}) \end{bmatrix},$$

$\mathbf{U}_i$ be a matrix with the conditional probabilities of responses $r_{i:k}$ given the cue profiles

$$\mathbf{U}_i = \begin{bmatrix} P(r_{i:1}|c_{1:1}, c_{2:1}) & P(r_{i:-1}|c_{1:1}, c_{2:1}) \\ P(r_{i:1}|c_{1:1}, c_{2:0}) & P(r_{i:-1}|c_{1:1}, c_{2:0}) \\ P(r_{i:1}|c_{1:0}, c_{2:1}) & P(r_{i:-1}|c_{1:0}, c_{2:1}) \\ P(r_{i:1}|c_{1:0}, c_{2:0}) & P(r_{i:-1}|c_{1:0}, c_{2:0}) \end{bmatrix},$$

and $\mathbf{V}$ be a matrix with the conditional probabilities of criterium values e_k given the cue profiles

$$\mathbf{V} = \begin{bmatrix} P(e_1|c_{1:1}, c_{2:1}) & P(e_{-1}|c_{1:1}, c_{2:1}) \\ P(e_1|c_{1:1}, c_{2:0}) & P(e_{-1}|c_{1:1}, c_{2:0}) \\ P(e_1|c_{1:0}, c_{2:1}) & P(e_{-1}|c_{1:0}, c_{2:1}) \\ P(e_1|c_{1:0}, c_{2:0}) & P(e_{-1}|c_{1:0}, c_{2:0}) \end{bmatrix}.$$

Individual i's achievement is then given as

$$A_i = \mathrm{tr}[\mathbf{U}_i' \mathrm{diag}(\boldsymbol{c})\mathbf{V}], \tag{5.3}$$

in which $\mathrm{tr}[\cdot]$ denotes the trace and $\mathrm{diag}(\boldsymbol{c})$ is the diagonal matrix with elements corresponding to $\boldsymbol{c}$.

The nonmetric lens model equation in (5.3) shows achievement is a function of three components: $\boldsymbol{c}$, which can be interpreted as the cue profile usability, $\mathbf{V}$, representing cue profile validity, and $\mathbf{U}_i$, representing cue profile utilisation. Maximum achievement is obtained by consistently giving as the response that event which has the highest conditional probability. Let o_m denote the optimal response for cue profile C_m. Then achievement is maximal if $P(r_i = o_m|C_m) = 1$ for all m. Consider an ecological system with parameters as given in Table 5.1. Highest achievement is obtained

Table 5.1: Ecological system with two dichotomous cues and a dichotomous criterion

C	$P(C)$	$P(C\|e_1)$	$P(C\|e_{-1})$	$P(e_1\|C)$	$P(e_{-1}\|C)$
$c_{1:1}$	.50	.80	.20	.80	.20
$c_{2:1}$	.40	.20	.60	.25	.75
$c_{1:1}, c_{2:1}$	.14	.16	.12	.57	.43
$c_{1:1}, c_{2:0}$	.36	.64	.08	.89	.11
$c_{1:0}, c_{2:1}$	.26	.04	.48	.08	.92
$c_{1:0}, c_{2:0}$	.24	.16	.32	.33	.67

by consistently giving response r_1 for cue profiles $C_1 = (c_{1:1}, c_{2:1})$ and $C_2 = (c_{1:1}, c_{2:0})$ and giving response r_2 otherwise. Thus, the optimal utilisation matrix is

$$\mathbf{U}^* = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

and the associated achievement is $A_i = .8$. Here, in a nutshell, lies the main difference between MCPL and NMCPL. In MCPL, achievement is maximal when the response system is identical to the ecological system. In NMCPL, however, achievement is maximised when the response system is different from the ecological system.

5.1.2 Cue validity and utilisation

The three components of achievement, $\boldsymbol{c}$, $\mathbf{V}$ and $\mathbf{U}_i$, are defined for cue profiles C_m, rather than for separate cues c_j. It has been customary to define validity and utilisation coefficients for separate cues. When all cues and event dimension are dichotomous, the phi-correlation between a cue and event (or response) can be taken as a validity (or utilisation) coefficient for that cue (Castellan, 1977). Edgell (1978, 1980, 1993) takes a different approach, in which not only the cues, but also cue profiles can be assigned a validity (or utilisation) coefficient. His approach amounts to partitioning the conditional probabilities $P(e_k|C_m)$ into several orthogonal components, representing the base-rate, main effects and interaction effects. While this approach works when the dichotomous cues are statistically independent with equal base-rates, it does not for a more general situation with dependent cues and/or unequal base rates. Moreover, neither phi-correlation nor the Edgell coefficients are suitable when cues and/or criterion are polytomous. In Appendix 5A, a method is proposed which is based on information theory (Shannon, 1948) and is applicable to the general situation of polytomous cues and criterion, where the cues may be dependent and have unequal base-rates. The validity coefficients defined in this manner reflect not only the conditional probability of the event given the cue, but also the cue base-rate. This is a nice property, since even though for instance $P(e_1|c_{1:1}) = 1$, so that the occurrence of the event is deterministically related to a cue, if $P(c_{1:1})$ is very small, the predictability of e on the basis of c_1 is not very high. As such, the validity of c_1

should not be considered high. For the environment as in Table 5.1, the validity of the base-rate is $\eta_0 = 0$, the validity of cue 1 is $\eta_1 = .28$, the validity of cue 2 is $\eta_2 = .13$ and the validity of the cue interaction is $\eta_{12} = -.04$. While the validity coefficients of the main effects can only be positive, sign is important for the validity coefficients of the cue interactions. When the validity is positive, it indicates that knowledge of the first (or second) cue increases the information of the second (or first) cue. When the coefficient is negative, as it is here, it indicates that knowledge of the first (or second) cue decreases the information of the second (or first) cue. Positive interaction validity coefficients indicate true interaction, while negative interaction validity coefficients indicate an information overlap (Attneave, 1959).

The information measures used do not only lead to generally applicable cue validity and utilisation coefficients. As noted in Appendix 5A, we may also define a predictability coefficient ψ for the total environment as the sum of all validity coefficients (for the ecological system in Table 5.1, $\psi = .36$). In similar fashion, the sum of the cue utilisation coefficients defines a consistency coefficient ξ_i for each individual.

5.2 Collaboration in NMCPL

When groups of individuals arrive at a single group response g_g, group achievement can be defined in an identical fashion to individual achievement, by replacing $\mathbf{U}_i$ in Equation 5.3 by $\mathbf{G}_g$, the $K \times M$ matrix containing the conditional probabilities of group responses $g_{g:k}$ given cue profiles C_m. When individuals form their own judgement before the collective judgement, the group cue utilisation matrix $\mathbf{G}_g$ will depend on the individual cue utilisation matrices $\mathbf{U}_i$. If the group functions under a simple majority rule, then each element g_{km} in $\mathbf{G}_g$ will represent the probability of a majority of the group giving a response k conditional on cue profile m. The individual responses of the group members are assumed to be conditionally independent, so that the probability of a majority giving a response for a cue profile C_m is completely specified by the conditional probabilities of the responses of the different group members. For instance, for a group consisting of three individuals

$$\begin{aligned} P(r_{g:1}|C_m) = {} & P(r_{1:1}|C_m)P(r_{2:1}|C_m)P(r_{3:1}|C_m) + P(r_{1:1}|C_m)P(r_{2:1}|C_m)P(r_{3:-1}|C_m) \\ & + (r_{1:1}|C_m)P(r_{2:-1}|C_m)P(r_{3:1}|C_m) + P(r_{1:-1}|C_m)P(r_{2:1}|C_m)P(r_{3:1}|C_m). \end{aligned}$$

Similar to individual achievement, group achievement is maximal when the group consistently gives responses with the highest conditional probability of being correct.

If individuals respond independently, the probability that the collective response is correct can be derived from the probability that an individual response is correct. We shall denote this latter probability as $p_i = P(r_i = e_k)$. If this probability is identical for all individuals, i.e. $p_i = p_j = p$, and individuals respond independently, the probability that a collective response by simple majority is correct in a group with an odd number n of individuals is

$$P(r_g = e_k) = \sum_{m=\frac{n+1}{2}}^{n} \binom{n}{m} p^m (1-p)^{(n-m)}. \tag{5.4}$$

This result was already derived in 1785 by Marquis de Condorcet, and his main result is now known as the Condorcet Jury Theorem (CJT):

Theorem 1 (Condorcet Jury Theorem).
If $p > .5$, $P(r_g = e_k)$ increases monotonically in n, while if $p < .5$, $P(r_g = e_k)$ decreases monotonically in n, and

$$\lim_{n\to\infty} P(r_g = e_k) = \begin{cases} 1 & \text{if } p > .5 \\ .5 & \text{if } p = .5 \\ 0 & \text{if } p < .5. \end{cases}$$

The CJT has received much attention over the years (Berend & Paroush, 1998; Boland, 1989; Kanazawa, 1998; T. Koch & Ridgley, 2000; Owen, Grofman, & Feld, 1989). One avenue of research, with particular relevance for the present study, has been to generalise the theorem to situations in which group members have different competency (e.g. Ladha, 1992; Owen et al., 1989). The generalised CJT is essentially similar to the original CJT, with p replaced by $\overline{p}$, the mean probability of a correct response in the group, with the exception that, if the distribution of p is asymmetric, $\lim_{n\to\infty} P(r_g = e_k) \neq .5$ when $\overline{p} = .5$ (Owen et al., 1989).

From the generalised CJT, it can be inferred that for groups functioning under a simple majority rule, the probability that the group response is optimal approaches 1 as group size increases, as long as the mean probability of an optimal individual response is greater than .5. Even though none of the individual cue utilisation matrices might be optimal, as group size increases, the group cue utilisation matrix can approach optimality. It is in this sense that individual bias can be corrected in a group decision.

The inference from the generalised CJT about the optimality of $\mathbf{G}_g$ holds only if group members base their responses on all available information. In the case that all group members base their responses on the same partial ecological system, the group utilisation matrix will approach the optimal utilisation matrix for this partial ecological system, but the optimal utilisation matrix for the complete ecological system may be different. For instance, the optimal response for the partial cue profile $C_m^1 = (c_{1:1}, c_{2:1})$ is r_1, while the optimal response for the complete cue profile $C_m = (c_{1:1}, c_{2:1}, c_{3:1}, c_{4:1}, c_{5:0}, c_{6:1})$ is r_{-1}. The situation under a division of labour is different. If the individual group members base their responses on different partial ecological systems, the group response matrix $\mathbf{G}_g$ will, besides differences in the individual utilisation of the cues, also reflect differences in the relations between cues and criterium. If the complete ecological system consists of six cues and an event, and the group consists of three members, each of which was trained in a different partial ecological system consisting of two cues and an event, the group as a whole can base its decisions on the complete ecological system. In such a situation, there will be no correction for individual bias, but there will be a correction for neglecting relevant evidence. The extent of this correction depends on the group process. Suppose the optimal response for the partial cue profile $C_m^1 = (c_{1:1}, c_{2:1})$ is r_1, the optimal response for the partial cue profile $C_m^2 = (c_{3:1}, c_{4:1})$ is r_1 and the optimal response the partial cue profile $C_m^3 = (c_{5:0}, c_{6:1})$ is r_{-1}. If the group members respond optimally to their

partial cue profile and the group decides by simple majority, then the group would respond $r_{g:1}$ to the complete cue profile $C_m = (c_{1:1}, c_{2:1}, c_{3:1}, c_{4:1}, c_{5:0}, c_{6:1})$, with the probability $P(r_{g:1})$ increasing as the consistency of the group members increases. But, as before, the optimal response to the complete cue profile is $r_{g:-1}$. As such, a simple majority rule may lead to sub-optimal group achievement, even though the group members all respond optimally to the information at hand.

5.2.1 Optimal group process

Situations such as the one just described call for a weighted majority rule. If each group member has a certain probability $p_i = P(e_k|C_m)$ that their response $r_{i:k}$ to cue profile C_m is correct, the optimal group decision rule is the weighted majority rule (see Appendix 5B):

$$r_g = \text{sgn}\left\{\sum_{i=1}^{n} \log\left(\frac{p_i}{1-p_i}\right) r_i\right\}. \tag{5.5}$$

To distinguish this process from others, it will be referred to as the weighting-by-evidence (WE) rule, since each response is weighted by its evidential support. For example, take an ecological system with three cues c_1, c_3, and c_5, that have identical properties to c_1 in Table 5.1, and three cues c_2, c_4, and c_6, that have identical properties to c_2 in Table 5.1. This is the ecological system used in experiment 1. All cues are conditionally independent given the events e_k. If group member 1 responds optimally to partial cue profile $C_m^1 = (c_{1:1}, c_{2:1})$, he would respond $r_{1:1}$ with a probability of $p_1 = .57$ that the response is correct. The same holds for group member 2 who responds optimally to partial cue profile $C_m^2 = (c_{3:1}, c_{4:1})$. If group member 3 responds optimally to partial cue profile $C_m^3 = (c_{5:0}, c_{6:1})$, he would respond $r_{3:-1}$ with a corresponding probability of $p_3 = .92$ that this response is correct. In this case, $\sum_i r_i \log\left(\frac{p_i}{1-p_i}\right) = -1.88$, so the weighted majority rule would result in $r_{g:-1}$. This is the optimal response, since $P(e_{-1}|c_{1:1}, c_{2:1}, c_{3:1}, c_{4:1}, c_{5:0}, c_{6:1}) = .87$. If member 3 were to give the suboptimal response $r_{3:1}$, he would have a corresponding probability of .08 that this response is correct. Even though the group is now unanimous in giving response $r_{i:1}$, the weighted majority rule would still result in $r_{g:-1}$, since $\sum_i r_i \log\left(\frac{p_i}{1-p_i}\right) = -1.88$, due to the lack of evidential support for e_1. In the completely distributed case, the optimal weighted majority rule *always* results in the optimal group response. This is due to the fact that, in the present ecological system, the cues are conditionally independent given the events, so that

$$\log\left(\frac{P(e_k|c_1,\ldots,c_6)}{1-P(e_k|c_1,\ldots,c_6)}\right) = \sum_j \log\left(\frac{P(e_k|c_j)}{1-P(e_k|c_j)}\right).$$

This means that the log-odds of the events for a complete cue profile is reconstructable from the log-odds of the events of the partial cue profiles. Hence, under the weighting-by-evidence rule, the group task is a divisible task.

While weighting-by-evidence is the optimal group process, it requires precise knowledge of the probabilities p_i of events given the cue profiles, which is an unrealistic assumption. However, it is not unrealistic that individuals have some indication of

the amount of correct responses they gave to partial cue profiles. If individuals know their conditional achievement

$$(A_i|C_m) = \sum_k P(r_{i:k}|C_m)P(e_k|C_m), \tag{5.6}$$

then a weighting-by-achievement process may be a viable alternative to weighting-by-evidence. Similarly to the latter, the weighting-by-achievement (WA) rule will be defined as

$$r_g = \operatorname{sgn}\left\{\sum_{i=1}^{n} \log\left(\frac{(A_i|C_m)}{1-(A_i|C_m)}\right) r_i\right\}, \tag{5.7}$$

As in the WE rule, weights are determined by log-odds, but now those of individual i giving a correct response to partial cue profile C_m^i, rather than i's particular response $r_{i:k}$ to C_m^i being correct. If individuals are completely consistent in their responses, so that $P(r_{i:k}|C_m) = 1$ for one k, and 0 otherwise, then achievement is identical to p_i, and (5.7) will lead to identical results as (5.5). If individuals are not entirely consistent, (5.7) will result in some process loss compared to (5.5). Consider the situation just described for the optimal weighted majority rule, and suppose that $P(r_{1:1}|c_{1:1}, c_{2:1}) = .8$, $P(r_{2:1}|c_{3:1}, c_{4:1}) = .8$, and $P(r_{3:1}|c_{5:0}, c_{6:1}) = .1$. In this case, the conditional achievements are $(A_1|c_{1:1}, c_{2:1}) = .54$, $(A_2|c_{3:1}, c_{4:1}) = .54$, and $(A_3|c_{5:0}, c_{6:1}) = .84$, respectively. For responses $r_{1:1}$, $r_{2:1}$, and $r_{3:-1}$, $\sum_i \log\left(\frac{(A_i|C_m)}{1-(A_i|C_m)}\right) r_i = -1.29$, so the group response would be $r_{g:-1}$. If individual 3 would respond $r_{3:1}$ however, the group response would be $r_{g:1}$, unlike that prescribed by the WE rule. However, $P(r_{g:1}|C_m)$, the probability that the group response to C_m is $r_{g:1}$, is in this case identical to $P(r_{1:1}, r_{2:1}, r_{3:1}) = .07$, so the probability that the group gives the sub-optimal response under the WA rule is not very high. For comparison, the probability of the group giving the sub-optimal response under a simple majority rule is $P(r_{g:1}|C_m) = .67$, which is clearly much higher.

5.2.2 Predicting group achievement

The group utilisation matrix $\mathbf{G}_g$ for the complete ecological system can be derived from the individual utilisation matrices of the partial ecological systems by making two assumptions. The first regards consistency of individual responses over partial and complete ecological systems. This consistency requires individuals to respond in the complete ecological system as if they observed only the partial ecological system in which they were trained. In other words, if individual i learned in a partial ecological system consisting of c_1 and c_2, then $P(r_i|c_{1:j}, c_{2:k}, c_{m:1}, \ldots) = P(r_i|c_{1:j}, c_{2:k}, c_{m:0}, \ldots)$ for all $m = 3, \ldots, 6$. If so, the individual utilisation matrix $\mathbf{U}_i$ for the complete ecological system is entirely determined by the individual cue utilisation matrix of the partial ecological system.

In order to derive the group utilisation matrices $\mathbf{G}_g$ from the individual utilisation matrices for the complete ecological system, an assumption must be made regarding the group process. A group process that seems plausible and which leads to particularly easy computation, is the simple majority (SM) process. In this case, we will write

$\mathbf{G}_{\text{sm}}$ for the predicted group utilisation matrix. All elements g_{km} in $\mathbf{G}_{\text{sm}}$ will correspond to the probability that a majority of group members responds with alternative k for cue profile m. For a weighting-by-achievement processes, computation of the expected group score is somewhat more complicated. We will write $\mathbf{G}_{\text{wa}}$ for the predicted group utilisation matrix. For the WA rule of (5.7) and dichotomous responses, elements g_{1m} in $\mathbf{G}_{\text{wa}}$ will now correspond to the probability $P(\sum_i \log(\frac{(A_i|C_m)}{1-(A_i|C_m)})r_i > 0)$, and elements g_{-1m} to the probability $P(\sum_i \log(\frac{(A_i|C_m)}{1-(A_i|C_m)})r_i < 0)$.

5.2.3 Collective gains and process loss in NMCPL

The two reasons why groups can outperform individuals when it comes to decision-making, mentioned at the beginning of this chapter, can be re-interpreted from the viewpoint of NMCPL. Individual bias can be defined in terms of the difference between individual responses and optimal responses. That is, a natural definition of (conditional) bias is $1 - P(r_i = o_m|C_m)$. From the generalised CJT, it follows that, if individual bias is not too extensive, so, on average, $P(r_i = o_m|C_m) > .5$, then $P(r_g = o_m|C_m) \rightarrow 1$ as group size increases and the group process is a simple majority process, or an adequate weighted majority process. Hence, group decisions are less affected by individual bias than individual decisions. For groups to realise the potential performance due to a larger informational base, the individual responses based on different information must be properly combined in the group response. As mentioned earlier, the optimal group process under a complete distribution of information is the weighting-by-evidence process. Weighting-by-achievement, which is a more plausible group process, will lead to some process loss. The extent of this process loss depends on the consistency of individuals. For entirely consistent individuals, there is no process loss, since weighting-by-evidence and weighting-by-achievement are identical in this case. The process loss when individuals are not entirely consistent will be smaller than the process loss associated with a simple majority process.

While weighting-by-achievement is a more plausible model of the group process than weighting-by-evidence, this is not to say that it gives an accurate description of the way in which groups actually arrive at a collective response. It is highly unlikely that group members report their conditional achievement for the given partial cue profiles, then take log-odds transforms of these and weight their responses accordingly in order to arrive at a group response. Weighting-by-achievement is not a model of the overt group process, but taken as a model of the underlying group process, a model of the influence each individual group member has in the collective response. It is not unlikely that influence in the group response is dependent on conditional achievement. The weighting-by-achievement process defined in (5.7) assumes that the relation between influence and conditional achievement is non-linear. In the actual group process, member influence will be more directly dependent on confidence, rather than conditional achievement. Confidence may be influenced by many factors, but it is assumed here that it has a strong relation with conditional achievement. Under this assumption, the group process based on confidence may be quite similar to a weighting-by-achievement process. Previous research has shown that in intellective tasks, confidence is positively related to accuracy, while in judgemental

tasks, this relation is usually not found (Zarnoth & Sniezek, 1997). NMCPL tasks are not truly intellective, since there is no demonstrably correct answer. Therefore, one might expect confidence to be unrelated to achievement in these tasks. However, in the judgemental tasks in which confidence was unrelated to accuracy, no feedback on the correct answer was given. In NMCPL, such feedback is an integral part of the task. By providing objective feedback on judgement accuracy, it is expected that this accuracy will be reflected in confidence. Zarnoth and Sniezek (1997) have shown that member confidence affects their influence in a final group judgement or decision. While the actual group process may be largely dependent on confidence, if confidence itself is largely dependent on conditional achievement, the group process may indirectly mimic the weighting-by-achievement process. The likeness of the two will depend on how weights determined by confidence resemble the weights prescribed by the weighting-by-achievement process. This resemblance itself will be dependent on the strength, and form, of the relation between confidence and achievement. In this way, there are at least two causes of process loss. The first is determined by the difference between the optimal group process (weighting-by-evidence) and the underlying group process (weighting-by-achievement), and the second is determined by the difference between the underlying group process (weighting-by-achievement) and the actual group process (weighting-by-confidence).

5.3 Experiment 1

The first experiment investigated the effect of different distributions of information over group members on group achievement. In one condition (the distributed condition), all information was completely distributed over group members, so that each group member possessed only unique information. In the second condition (the shared condition), all information was shared among group members, so that all group members possessed the same information. When the information is completely distributed, the group as a whole possesses more information than any individual alone, so group resources are larger than individual resources. For this reason, the group has a higher potential achievement than any individual. Realisation of this potential requires an adequate group process in which the available information is properly integrated in the group response. When information is completely distributed, the optimal group process is the weighting-by-evidence rule. Weighting-by-achievement will lead to some process loss, depending on the consistency of group members. When all information is shared, the group resources are the same as individual resources, and the group does not have a higher potential achievement than individual group members. However, when individual achievement lies below potential individual achievement, because individuals do not consistently give optimal responses, the group may still be advantaged over individual members. When the group process is a simple majority process, this advantage follows from the generalised CJT. A simple majority process is not a prerequisite; a weighted majority process such as weighting-by-achievement will also result in higher group achievement.

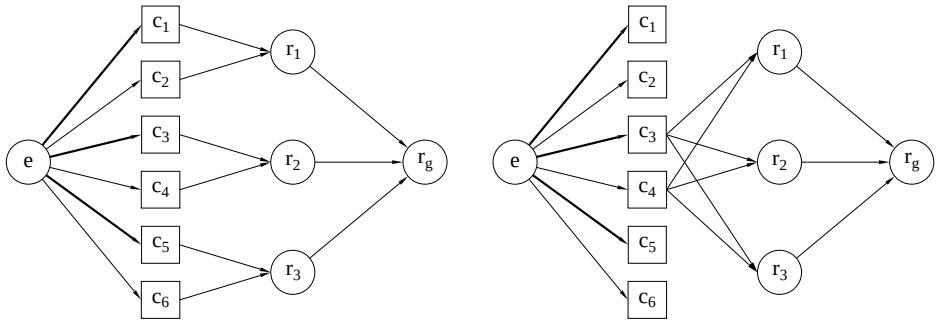
In the first experiment, participants were individually trained in a partial ecological system. This individual task was followed by a group task, in which groups were

to make decisions in the complete ecological system. In the distributed condition, the partial ecological systems in which individual group members were trained as a whole covered the complete ecological system. In the shared condition, the partial ecological systems in which the group members were trained were identical, so the group as a whole had knowledge about the same partial ecological system as any individual group member. As such, groups in the distributed condition were advantaged over groups in the shared condition. However, full realisation of potential group achievement in the distributed condition required the weighting-by-evidence process, while a simple majority process could suffice for the realisation of potential performance in the shared condition. The distribution of information over group members, and the realisation that some (partial) cue profiles have a higher predictive validity than others, was expected to move groups in the distributed condition to a weighting-by-achievement process. When all information is shared, there is less impetus for such a weighted majority process. Since group members base their responses on the identical information in the shared condition, much more initial agreement is expected in this condition than in the distributed condition, where group members base their responses on different information. The more members agree, the less reason they have for assuming differences in their predictive accuracy. As such, there is less reason to weight members' contributions differently. On the other hand, even if individuals base their responses on the same information, reasonably large differences in achievement are possible due to differences in consistency. In those cases, there is good reason to weight members contributions differently. For this reason, there is no clear expectation regarding the group process in the shared condition. Regardless of the group process in the shared condition, if groups in the distributed condition function under a weighting-by-achievement process, as expected, their performance should be higher than that of groups in the shared condition. This is the first hypothesis that will be tested in the experiment. Besides testing this hypothesis, the overall purpose of the first experiment is to determine the underlying group process and to see whether it depends on the distribution of information over group members. Investigation of the group process will be based on a comparison between the group responses and predictions derived from different group processes, as well as on the effect of confidence on the group responses. As indicated, the expected group process in the distributed condition is the weighting-by-achievement process. This is the second hypothesis to be tested.

5.3.1 Method

Participants and design

Ninety university undergraduates participated in the experiment. There were 25 males and 65 females. The mean age was 22.02 ($SD = 1.73$). The experiment had two conditions. In the shared condition, all group members learned to predict an event on the basis of the same two cues. In the distributed condition, all group members learned to predict an event on the basis of two different cues. Since the complete ecological environment consisted of six cues, groups in the shared condition had information about only a partial ecological system, while groups in the distributed



(a) Distributed (b) Shared

Figure 5.1: Group task systems in experiment 1. Thick arrows between e and c_j indicate relatively high validity ($\eta_j = .28$) and thin arrows relatively low validity ($\eta_j = .13$) of the cues.

condition had information about the complete ecological system.

Task

The NMCPL task was a medical diagnosis task, in which participants were to decide whether a patient has disease A ('Atoritus') or B ('Burtosis') on the basis of the presence or absence of a number of symptoms. These symptoms were labelled 'diminished appetite', 'dizziness', 'fever', 'headache', 'nausea' and 'tiredness'.

The total ecological system consisted of 6 dichotomous cues and one dichotomous criterion. The criterion had a base rate of $P(e_1) = P(e_{-1}) = .5$. The cues had similar characteristics as those in Table 5.1, with cues c_1, c_3 and c_5 identical to c_1 in Table 5.1 and cues c_2, c_4 and c_6 identical to c_2 in Table 5.1. In the individual task, only two cues were presented to each participant, either the pair (c_1, c_2), (c_3, c_4) or (c_5, c_6). Each participant thus based their responses on a partial ecological system. In the shared condition, the partial ecological system was the same for all group members. In the distributed condition, the partial ecological system was different for each group member. In the group task, the whole ecological system was presented. However, it was expected that the individuals would base their individual responses mainly on the partial ecological system they encountered in the individual task. A graphical representation of the difference between the conditions is given in Figure 5.1. The probabilities defining the total ecological system are given in Appendix 5C.

The maximum potential achievement for the complete ecological system was .92. This was also the maximum potential achievement of groups in the distributed condition if they were to function under a weighting-by-evidence process. For groups in the shared condition, the maximum potential achievement was .80 under a weighting-by-evidence process. The larger group resources in the distributed condition clearly affected the potential group achievement. In the group trials, not all possible 64 cue

profiles were given. The group task consisted of 20 trials and the 20 corresponding cue profiles were chosen so that, if groups in both conditions would work under a simple majority process, their performance would be reasonable close. More in particular, a number of cue profiles were included for which, if all group members would respond optimally to their partial cue profiles, a majority of group members would respond sub-optimally to the complete cue profile. For this reason, groups in the distributed condition would need to use a weighted majority – which would then favour the minority position – rather than a simple majority process, in order to realise the potential group achievement.

For the chosen cue profiles, maximum potential achievement was .73. Groups in the distributed condition, functioning under the weighting-by-evidence process, would be able to reach this score. Groups in the shared condition had less information, and hence their potential performance was lower. However, realisation of this potential was less dependent on the group process. If all group members would respond optimally to the information, and groups functioned under either a simple majority, weighting-by-evidence or weighting-by-achievement process, groups in this condition would, on average, reach a maximum score of .62. For groups in the distributed condition, functioning under a simple majority process with all members responding optimally to their information, achievement would be .66.

Procedure

Participants entered the experimentation room in groups of three and were each placed in front of a computer. The computers were placed in such a way that participants could not see each other's screens, but, when looking up, could see each other's faces in order to allow for group discussion in the group task. In the instructions, participants were informed they were to diagnose the presence of either one of two diseases in a number of patients. The diseases were linked to six symptoms, but in order to facilitate learning of the relations between symptoms and diseases, each participant would be shown information about only two of these symptoms. Participants were made aware the relation between symptoms and disease was probabilistic, and that it would not be possible to always make the correct diagnosis. After reading the instructions, participants could proceed with the individual task, consisting of a total of 100 trials. The order of the trials was randomised for each participant. In each trial, the values of two cues were presented by the cue label (the name of the symptom) and a '+' or '−' beside it, to indicate presence or absence of the symptom respectively. The cue values were presented below each other, and the order was randomised for each trial. Following the information, participants made a choice for either disease A or B. Immediately following this choice, participants received outcome feedback (i.e. they were informed about the true disease). Also, in the right-upper corner of the screen, participants received feedback on the total number of correct diagnoses they gave in the previous trials. Although it has not been shown that this last type of feedback enhances performance, contrary to other possible types of feedback, it does not have a detrimental effect on performance either (Castellan, 1974). The individual task was self-paced. After all group members completed the individual task, participants received instructions for the group task. In these instructions they

were informed they would now receive information regarding all symptoms, and were to arrive at collective diagnoses by discussing the case in the group. No further instructions were given as to how they should arrive at a collective response, so that groups were free in adopting a group process. After reading the instructions, groups proceeded with the group task consisting of 20 trials in random order. In each trial, the values of the six symptoms were presented to the participants (again in random order) on their computer screen, and were first asked to make an individual diagnosis. Following the diagnosis, participants were asked to rate their confidence on a nine-point scale, ranging from 'completely unconfident' to 'completely confident'. Then, they were instructed to discuss with the others in order to arrive at a group diagnosis. After reaching a group diagnosis, each participant indicated the group response on his computer and was then asked to indicate his confidence in the group diagnosis. The group task was followed by a computerised exit-interview.

Dependent measures

Obvious dependent measures are the individual achievement A_i and group achievement A_g. However, interest was not so much in the actually obtained achievement, but in the expected achievement for the given environment. While A_i and A_g are unbiased estimates of expected achievement, better estimates are given by what will be termed the individual and group scores. These scores are based on the probability that each response is correct, rather than whether each response is actually correct. The individual score S_i of individual i is defined as the mean probability of the response $r_{it:k}$ at trial t being identical to the event e over all individual trials T:

$$S_i = \frac{1}{T} \sum_{t=1}^{T} P(r_{it:k} = e). \tag{5.8}$$

Both observed achievement (proportion of correct responses) and the score defined above are unbiased estimates of the expected achievement of an individual in the given environment. But since the conditional probabilities $P(e_k|C_m)$ are known, they need not enter the estimation. Weighting the responses $r_{i:k}$ to C_m by $P(e_k|C_m)$, as is done in (5.8), gives an estimate with a smaller variance than the proportion $\widehat{p}(r_i = e)$ (see Appendix 5D for a proof).

For the ecological system of the individual task, as specified in Table 5.1, the theoretical minimum of the individual score was .2, which would be obtained by consistently giving the suboptimal response to each cue profile C_m. The theoretical maximum of the individual scores was .8, which would be obtained by consistently giving the optimal response to each cue profile C_m. In other words, the possible individual scores S_i lay in the closed interval $[.2; .8]$.

In a similar way to S_i, the group score S_g of group g is defined as the mean probability of the group response $r_{gt:k}$ at trial t being identical to the event e over all group trials T. For the group scores, the total environment was used (so that C_m had 64 levels, instead of 4 in the individual task). For the 20 cue profiles used in the group trials, the possible group scores lay in the closed interval $[.27; .73]$.

Both S_i and S_g are bounded from below and above. Hence, S_i and S_g will not follow a normal distribution, which is problematic for hypothesis testing by either F

or t-test. Hence, for hypothesis testing, the variables S_i and S_g were transformed by the following logit-transformation:

$$S' = \log\left(\frac{S - \min_S^*}{\max_S^* - S}\right),$$

in which $\min_S^* = \min_S - \varepsilon$ and $\max_S^* = \max_S + \varepsilon$ (the correction $\varepsilon = .01$ was applied in order to avoid problems when the score lies on the boundary).

Besides the individual and group scores, a final dependent measure was Conf, participants' reported confidence in their individual responses in the group task.

5.3.2 Results

The structure of this section is as follows. First, individual performance and cue utilisation in the individual task will be discussed. This is followed by the results related to group achievement. The hypothesis that groups in the distributed condition outperform groups in the shared condition is tested by comparing the group scores between these conditions. The remainder of the section will concern the group process. Three methods of analysis will be used. The first compares the obtained group scores to those expected from a simple majority or weighting-by-achievement process. The second method compares predictions of the two processes on the level of group responses, rather than overall scores. The third method investigates the relation between confidence and achievement on the one hand, and confidence and influence in the collective response on the other. Insofar as confidence is related to achievement and groups weight members' contributions according to confidence, this analysis, together with the previous ones, can provide evidence for the underlying group process.

Individual performance

In order to determine learning effects, the 100 trials in the individual task were divided into four blocks of 25 trials each. A repeated measures analysis for the individual scores showed a significant effect of Block, $F(3, 267) = 7.75$, $p < .001$. Pairwise contrasts, in which the individual score at each block was compared to the score at the previous block, showed that only the score at block 2 differed from the score at block 1, $F(1, 89) = 16.49$, $p < .001$. There was no significant difference between scores at later blocks. Since participants' performance did not noticeably change after the first 50 trials, participants' scores on the last 50 trials were used in the remaining analyses. The individual scores in the last 50 trials ranged from .38 to .80, with a mean of .68 ($SD = .10$). A t-test was performed to rule out initial differences between the conditions for the individual scores, since these may have affected later results. This test showed no difference between the individual scores in the distributed condition ($M = .68$, $SD = .09$) and the shared condition ($M = .68$, $SD = .11$), $t(88) = -.08$, $p = .934$.

Cue utilisation coefficients were computed for the responses in the last 50 trials. The mean of υ_0, the utilisation coefficient for the base-rate, was .01 ($SD = .02$). Such a low value indicates that participants did not show a preference for one of

the response categories. The mean of υ_1, the utilisation coefficient of cue 1, was .28 ($SD = .25$), and the mean of υ_2, the cue utilisation coefficient for cue 2, was .14 ($SD = .14$). The mean of υ_{12}, the utilisation coefficient for the cue interaction, was .02 ($SD = .07$). The value of this last coefficient indicates that, at least on average, participants based their responses on the separate cues rather than on the cue profiles. When the utilisation coefficients are compared to the validity coefficients, $\eta_0 = 0$, $\eta_1 = .28$, $\eta_2 = .13$ and $\eta_{12} = -.04$, we see that, on average, they are rather alike. Note that similarity of validity and utilisation coefficients is not indicative of high achievement. Maximum achievement could be obtained by consistently giving the optimal response to each cue profile. In the partial ecological system presented in the individual task, optimal responses were entirely determined by the value of c_1. Hence, for optimally responding participants, $\upsilon_1 = 1$. There was one participant with this value for υ_1. Overall, the cue utilisation coefficients indicated that participants did not consistently give optimal responses. This can also be seen in the consistency coefficient ξ, which had a mean value of .44 ($SD = .26$). Instead of maximising, participants' response behaviour was more indicative of probability matching, which is a response strategy that is often found in probability learning research (Castellan, 1977; Shanks, Tunney, & McCarthy, 2002). If participants were exactly matching probabilities, so that $\mathbf{U}_i = \mathbf{V}$, the expected achievement is $\mathrm{tr}[\mathbf{V}'\mathrm{diag}(\boldsymbol{c})\mathbf{V}] = .72$. While the mean individual scores were somewhat lower, they did not differ significantly from this last value, $t(89) = -1.56$, $p = .122$.

Group achievement

To test the hypothesis that groups in the distributed condition outperform groups in the shared condition, the group scores S_g were compared between the conditions. The mean group score in the distributed condition was .62 ($SD = .06$), while the mean group score was .58 ($SD = .05$) in the shared condition. A one sided t-test showed this to be a significant difference, $t(28) = 1.93$ $p = .032$. The first hypothesis was thus confirmed.

While groups in the distributed condition did perform better than those in the shared condition, the mean performance lay well below the maximum potential performance of .73. The mean group achievement in the shared condition was less removed from the mean potential performance of .62. If the group process was weighting-by-achievement, rather than weighting-by-evidence, realisation of the maximum potential performance required maximum individual performance of the group members (in which case weighting-by-achievement has identical results to weighting-by-evidence). As the results of the individual task indicate, individual performance was not maximal. Hence, if groups functioned under the weighting-by-achievement process, group achievement would still be below the maximal group achievement. With the procedure described in section 5.2.2, group achievement under different group processes could be predicted without requiring individuals to respond optimally to their partial cue profiles. As indicated in that section, derivation of the group utilisation matrices $\mathbf{G}_g$ from the individual utilisation matrices is based on the assumption that the individuals responded in the group task as they did in the individual task. This amounts to participants basing their response on only that information which they encountered

in the individual task (and thus ignoring four of the six symptoms they encounter in the group task), and using this information in the same way as in the individual task. This assumption was tested by a χ^2-difference test for each participant. This test compared the fit of two logistic regression models for the individual responses in the last 50 trials of the individual task and the 20 trials of the group task taken together. The first model had a single predictor with four levels, indicating the cue profile on which the response was been based. The second model included an additional interaction between cue profile and task (individual or group). If the responses to the cue profiles differed between the individual and group task, the second model should have a significantly better fit. Since there were a total of 90 tests, a significance level of $\alpha = .005$ was used for each test[1]. The tests showed that for 4 participants the hypothesis of equal responses in the individual and group task was untenable. This number was deemed sufficiently small, and hence the computation of the expected group scores should be valid.

For a simple majority (SM) process, the mean of the expected group scores was .61 ($SD = .03$) in the distributed condition and .59 ($SD = .04$) in the shared condition, which is a non-significant difference, $t(28) = 1.19$, $p = .25$. So, if groups in both conditions functioned under simple majority process, no significant difference in the group scores was expected. The observed group scores were compared to the expected group scores in a 2 (Condition) $\times$ 2 (Score: observed or expected) ANOVA with the last factor treated as a within-groups factor. There was no significant effect of this last factor, $F(1, 28) = .93$, $p = .343$, which indicates that the observed group scores were not significantly different than the expected group scores from a SM process. There was also no significant interaction between Condition and Score, so that the difference between observed and expected group scores was about equal for both conditions.

For a weighting-by-achievement (WA) process, the expected group scores were .65 ($SD = .04$) in the distributed condition, and .61 ($SD = .05$) in the shared condition. This is a significant difference, $t(28) = 2.29$, $p = .030$. Hence, if the group process was a WA process, a significant difference in group scores was to be expected. The observed group scores were compared to the expected group scores in a similar ANOVA as above. The effect of the Score factor was significant, $F(1, 28) = 6.59$, $p = .016$, which indicates that the observed group scores were significantly lower than the expected group scores from a WA process. There was no significant interaction between Condition and Score, which indicates that the difference between observed and expected group scores was about equal for both conditions. In both conditions, this difference was .03, indicating a small amount of process loss.

Group process

The analyses reported above indicate that both in the distributed and shared conditions performance lay below that predicted from a weighting-by-achievement process. However, comparisons between group achievement and expected group achievement under different group processes are not the best basis for making inferences about the group process. One sign of a weighted majority rule is the presence of group

[1]Bonferroni correction with $\alpha = .05$ would actually result in $\alpha' = \alpha/90 = .0005$, but it was thought that this value would reduce the power of the individual tests too much.

responses which follow a minority position. In the distributed condition, there were 81 unanimous group diagnoses, 201 by majority and 18 by minority. In the shared condition, there were 166 unanimous group diagnoses, 126 diagnoses by majority, and 8 diagnoses by minority. The higher frequency of unanimous group diagnoses in the shared condition was expected, since group members in this condition based their responses on the same partial cue profiles. More importantly, and also as expected, there were more minority decisions in the distributed as compared to the shared condition. Counting unanimous decisions as majority decisions, there were 282 majority and 18 minority decisions in the distributed condition and 292 majority and 8 minority decisions in the shared condition, which is a marginally significant difference, $\chi^2(1) = 3.26$, $p = .071$. The number of minority decisions was not very high in either condition, though. But a weighted majority process should result in a minority decision only if the weight of the minority member is larger than the combined weight of both majority members. In other words, the difference between minority and majority weight has to be quite substantial in order for a weighted majority process to result in a minority decision. When this difference is small in general, it will be difficult to infer the actual process from the frequency of minority decisions.

A more direct approach to inferring the group process is to compare group responses on each trial to the expected group response under different group processes. Since the derived group response matrices $\mathbf{G}_{\text{sm}}$ and $\mathbf{G}_{\text{wa}}$ contain probabilities of group responses for each cue profile C_m, they can be used to determine the likelihood of the group responses under each group process. One problem is that $\mathbf{G}_{\text{wa}}$ can contain deterministic predictions, i.e. $\mathbf{G}_{\text{wa}}$ may contain a number of 0's and 1's. Hence, the likelihood of group responses will be 0 if at least one group response occurs which was assigned a probability of 0. While these deterministic predictions could be corrected, a different approach was adopted here. The predictions were compared by the Root Mean Squared Error of Prediction (RMSEP), defined as

$$RMSEP = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[z_t - P(g_{gt:1})]^2}, \tag{5.9}$$

in which T is the total number of trials for each group, and $z_t = 1$ if the group response g_{gt} of group g at trial t is 1, and $z_t = 0$ for $g_{gt:-1}$. Each group was classified under the process with the lowest associated RMSEP. However, since none of the group processes may have been very accurate, the RMSEP's were also compared to the RMSEP of a null-model, which assigned each group response a probability of .5. The RMSEP of this null-model was .50. Using this procedure, 5 groups in the distributed condition could not be classified (the RMSEP of the two processes was higher than that of the null-model), 8 groups were classified as a simple majority process, and 2 as a weighting-by-achievement process. In the shared condition, 1 group could not be classified, 10 groups were classified as a simple majority process, and 4 groups as a weighting-by-achievement process. The relatively large number of unclassified groups in the distributed condition complicates the comparison between the conditions on the inferred group process. As yet, there is no clear evidence the difference in information distribution between the conditions has an effect on the group process.

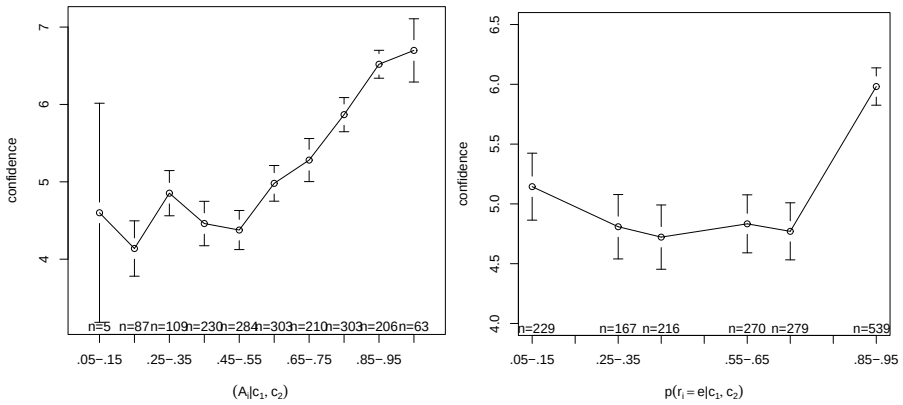


Figure 5.2: Relation between confidence ratings and $(A_i|c_1, c_2)$ and $P(r = e|c_1, c_2)$

Group process and confidence

The analysis above is a valuable means for making inferences of the underlying group process. But, as mentioned earlier, confidence was expected to play a more direct role in the actual group process than achievement. Achievement was expected to have an indirect effect through confidence. With that in mind, it was expected that confidence would be related to conditional achievement. Since it was not expected that the relation between achievement and confidence would necessarily be linear, Kendall's τ was used to investigate the strength of the relation. The rank-correlation between Conf and $(A_i|C_m^i)$ was $\tau = .27$ ($Z = 16.84$, $p < .001$). For comparison, the rank-correlation between Conf and $P(e_k|C_m^i)$, the probability that the actual response $r_{i:k}$ to partial cue profile C_m^i was correct, was $\tau = .14$, ($z = 8.93$, $p < .001$). The relation between the confidence ratings and both $(A_i|C_m^i)$ and $P(e_k|C_m^i)$ is depicted in Figure 5.2.

If responses were weighted by confidence, then the confidence of the minority group member should have been higher than the confidence of the majority group members when the group response followed the minority individual response. Conversely, when the group response followed the majority response, the confidence of the majority should have been higher than that of the minority. Table 5.2 contains the means and standard deviations of Conf for the minority member and majority members, for those group decisions in which there was initial disagreement ($n = 353$). As this table shows, minority confidence was indeed larger than majority confidence when the group adopted the minority response, while there was no clear difference in confidence when the group adopted the majority response. A 2 (Decision Type) $\times$ 2 (Condition) $\times$ 2 (Source) ANOVA was performed, in which the last factor represented the source of the confidence rating (minority or majority), and weighted least squares (WLS) estimation with weights $w = 1/\widehat{\sigma}^2_{\text{within}}$ was used to correct for unequal variances. This test showed significant main effects of Condition, $F(1, 1051) = 4.27$, $p = .039$, Decision Type, $F(1, 1051) = 4.62$, $p = .032$, and Source, $F(1, 1051) = 3.90$, $p = .049$. More importantly, there was a significant interaction between Decision Type and

Table 5.2: Minority and majority confidence ratings by condition and decision type

	Distributed				Shared			
	Majority		Minority		Majority		Minority	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Conf minority	5.26	1.85	6.72	1.74	4.19	2.07	5.50	2.45
Conf majority	5.13	2.10	4.61	2.17	4.74	2.20	4.94	1.65

Source, $F(1, 1051) = 7.36$, $p = .007$. Two post-hoc t-tests, with a Welch-correction for the degrees of freedom, showed that minority confidence was significantly higher than majority confidence for minority decisions, $t(49.98) = 3.37$, $p = .001$, while there was no significant difference between minority and majority confidence for majority decisions, $t(693.18) = -0.96$, $p = .339$.

To directly test how member confidence influenced the group response, three logistic regression models were compared. The first model represented a simple majority process, and had the three individual responses $r_i \in (-1, 1)$ as predictors:

$$M_1 : \log\left(\frac{P(r_{g:1})}{1 - P(r_{g:1})}\right) = \alpha + \sum_{i=1}^{3} \beta_i r_i. \tag{5.10}$$

In order to specify the model for a weighted majority process with weights determined by confidence, the functional form of the relation between confidence and weight should be specified. This functional form may not only be dependent on the relation between confidence and influence in the group response, but also on how the response scale was used to indicate actual confidence. Two functional forms were deemed plausible: linear and exponential. The model representing a weighted majority process with the weights linearly dependent on member confidence was

$$M_2 : \log\left(\frac{P(r_{g:1})}{1 - P(r_{g:1})}\right) = \alpha + \sum_{i=1}^{3} \beta_i \frac{r_i x_i}{\max(x)}, \tag{5.11}$$

in which x_i is i's confidence rating in response r_i, and $\max(x) = 9$. The division by $\max(x)$ was for scaling purposes, so that the estimated regression weights β_i were comparable to those of M_1. The model representing a weighted majority process with weights being an exponential function of confidence was

$$M_3 : \log\left(\frac{P(r_{g:1})}{1 - P(r_{g:1})}\right) = \alpha + \sum_{i=1}^{3} \beta_i r_i e^{\frac{x_i - \overline{x}}{\max(x)}}, \tag{5.12}$$

in which $\overline{x} = 5.19$. The results for the three logistic regression models are given in Table 5.3. The fit measure $-2 \log L$ represents the log likelihood ratio of the fitted model against the saturated model (e.g. Dobson, 2002). Since all three models had four parameters and were estimated with the same number of observations, selecting between the models on the basis of this measure is equivalent to selecting on the

Table 5.3: Three logistic regression models for $g_{g:1}$

	M_1			M_2			M_3		
	est.	Z	p	est.	Z	p	est.	Z	p
α	0.43	2.01	.04	0.18	1.05	.30	0.34	1.61	.11
β_1	2.82	9.95	$< .001$	5.03	8.79	$< .001$	3.16	9.44	$< .001$
β_2	2.51	10.29	$< .001$	4.62	9.34	$< .001$	2.80	9.83	$< .001$
β_3	2.54	9.50	$< .001$	4.68	8.80	$< .001$	2.84	9.27	$< .001$
$-2 \log L$	179.57			219.95			170.00		
				Distributed					
α	0.61	2.22	.03	0.26	1.19	.24	0.50	1.83	.07
β	2.53	9.35	$< .001$	4.91	7.81	$< .001$	2.90	8.59	$< .001$
$-2 \log L$	119.05			133.16			108.30		
				Shared					
α	0.10	0.26	.79	0.04	0.16	.88	0.06	0.19	.85
β	2.77	7.72	$< .001$	4.41	6.50	$< .001$	2.88	6.93	$< .001$
$-2 \log L$	60.62			87.54			63.30		

basis of either C_{AIC} or C_{BIC} (see Chapter 3). As can be seen in this table, M_3 fitted best and was selected over both M_1 and M_2. Hence, it should be concluded that overall, groups functioned under a weighted majority process with weights being an exponential function of confidence. In order to test hypothesis for differences in the group process between conditions, the analysis above was repeated for each condition separately. Since the estimated values of the β_i parameters were similar, and no difference between group members was actually expected, the three models were reduced to 2-parameter models, with the single predictor now defined as the sum of the original predictors. Parameter estimates and model fits for each condition are given in the lower part of Table 5.3. As expected, in the distributed condition, M_3 was selected over M_1 and M_2, while in the shared condition, M_1 was selected over M_2 and M_3. Taking confidence into account, there is support for expectation that groups in the distributed condition functioned under a weighted majority rule. There was no strong evidence for this process in the shared condition. As such, the distribution of information over group members did appear to affect the group process.

5.3.3 Discussion

The first hypothesis was confirmed: group achievement was higher in the distributed condition than in the shared condition. Regarding the expected group process in the distributed condition, results were somewhat inconclusive. If groups in both conditions functioned under a simple majority process, then no differences in group achievement were expected. The difference in group achievement between the conditions indicates that it is unlikely that groups in both conditions functioned under a simple majority process. The actual performance of groups in the distributed condition lay between the expected performance from a simple majority process and the expected performance from a weighting-by-achievement process. Actual group achievement in the shared condition was slightly below the expected performance from a simple majority process, and further removed from the expected performance associated with a weighting-by-achievement process. The comparison of group achievement

to expected performance leads to the conclusion that groups in the shared condition mainly functioned under a simple majority process, while some groups in the distributed condition functioned under a simple majority process, and some under a weighting-by-achievement process. Inference of the group process from comparisons between the group responses (rather than group achievement) and predictions about these derived from different group processes showed that in both conditions, the number of groups that appeared to function under a weighting-by-achievement process was much smaller than the number of groups that appeared to function under a simple majority process. Hence, the expectation that groups in the distributed condition would mainly function under a weighting-by-achievement process was not confirmed in this analysis. The weighting-by-achievement process as formalised in (5.7) should be thought of as an idealised model of individual influence in the collective decision. A more accurate representation of the actual group process should take individual confidence into account. As expected, confidence was related to conditional achievement, so that a weighting-by-confidence process can be taken as an indirect version of the weighting-by-achievement process. While there was evidence that groups in the distributed condition weighted members contributions according to their confidence, there was no evidence for such a process in the shared condition. Hence, when focussing on what is thought to be a more accurate representation of the actual group process, the expectation that groups in the distributed condition mainly functioned under a weighted majority process was confirmed.

The finding that groups in the shared condition mainly functioned under a simple majority process should be treated with some caution. Due to the smaller frequency of minority decisions in this condition, a model in which individual contributions are weighted by confidence has less room to provide a better fit to the data than a model in which contributions are not weighted. It may be that groups are inclined to use the same group procedure, regardless of how information is distributed over group members. Insofar as the natural group process is weighting-by-confidence, which indirectly represents a weighting-by-achievement due to the relation between confidence and achievement, there are multiple causes of process loss. The first concerns process loss due to weighting-by-achievement rather than weighting-by-evidence. The second is due to the imperfect relation between confidence and achievement. From the predicted group scores, it is clear that the bulk of process loss should be attributed to the first cause. Compared to the expected group achievement from the weighting-by-achievement process, groups in both conditions exhibited a small amount of process loss.

5.4 Experiment 2

The purpose of the second experiment was to study collaboration in NMCPL under a different distribution of information over group members. As mentioned in the introduction, research has shown that groups tend to focus on information that is shared, rather than the unique information group members can contribute. One explanation of this common knowledge effect (Gigone & Hastie, 1993) is based on an information sampling model (Stasser & Titus, 1985; Larson et al., 1996). This

model is based on the idea that each piece of information has a fixed probability of being mentioned by a group member in the group discussion. As more members share a piece of information, the probability that this information will be mentioned is larger. A problem with previous research on the common knowledge effect is that the information provided to the group members is unfamiliar to them. A notable exception is the study by Larson et al. (1996), but while they found evidence for the common knowledge effect, they failed to investigate how it affects the collective decision itself. It is this last effect which is of interest here. While the information sampling model can explain why shared information is mentioned more often and earlier in group discussions, it does not explain why shared information has greater impact on the collective decision even when unshared information is mentioned. A plausible explanation for this effect is in terms of social validation (see Chapter 4), where the validity of unfamiliar information is inferred from the number of individuals who share the same piece of information. If this is an adequate explanation, the effect should not be found when individuals have other means of inferring the validity of their information. More precisely, if the validity of shared information is high relative to the validity of unique information, the shared information will have a large impact on the collective decision. If the validity of the shared information is low relative to the validity of the unique information, the impact of unique information on the collective decisions should be larger than that of the shared information. In the second experiment, the distribution of information over group members was such that information was partly shared, partly unique. In one condition, the validity of the shared information was higher than that of the unique information, and in the other condition, the validity of the shared information was relatively low compared to that of the unique information. As such, the common knowledge effect is expected in the first condition, but not in the second.

The situation of partly shared, partly unique information was not studied in experiment 1. In this case, the optimal group process is more complicated. Unlike the situation with completely distributed information, the log-odds for the complete cue profiles are not reconstructable from the log-odds of the partial cue profiles. For example, take an ecological system consisting of four cues, c_1 like c_2 in Table 5.1 and c_2, c_3 and c_4 like c_1 in Table 5.1. Cue 1 is shared by all group members, while the other cues are unique. This is the ecological system for the low validity conditions used in this experiment. If group member 1 responds optimally to partial cue profile $C_m^1 = (c_{1:0}, c_{2:0})$, he would respond $r_{1:-1}$ with a probability of $p_1 = .67$ that the response is correct. The same holds for group member 2 who responds optimally to partial cue profile $C_m^2 = (c_{1:0}, c_{3:0})$. If group member 3 responds optimally to partial cue profile $C_m^3 = (c_{1:0}, c_{4:1})$, he would respond $r_{3:1}$ with a corresponding probability of $p_3 = .89$ that this response is correct. Under the weighting-by-evidence rule, the group response would be $g_{g:1}$, since $\sum r_i \log\left(\frac{p_i}{1-p_i}\right) = .67$. However, for the complete cue profile, $P(e_1|c_{1:0}, c_{2:0}, c_{3:0}, c_{4:1}) = .33$, so the optimal response is in fact $g_{g:-1}$. The problem is that the weighting-by-evidence rule in a situation of partly shared, partly unique information, gives too much weight to the shared cue. For the weighting-by-achievement rule, and group members responding by probability matching, the conditional achievements are $A_1 = .56$, $A_2 = .56$ and $A_3 = .80$

and $\sum r_i \log\left(\frac{(A_i|C)}{1-(A_i|C)}\right) = .95$. Hence, the group response will also be sub-optimal. However, while $P(r_{g:1}) = 1$ under the weighting-by-evidence rule, so it always results in the suboptimal answer, $P(r_{g:1}) = .11$ under the weighting-by-achievement rule. In this respect, the weighting-by-achievement rule is actually better than the weighting-by-evidence rule, which is the optimal rule when the information is completely distributed. For a final comparison, if group members respond by probability matching, then we would have $P(r_{g:1}) = .50$ under a simple majority rule.

A difference between this experiment and the first one was the inclusion of conditions in which objective feedback was given in the group trials. One reason why confidence may determine the members' influence in the collective decision is that it is the only indication of the validity of members' decisions. Objective feedback provides other evidence for members' accuracy, and may weaken the effect of confidence on members' influence in the collective decision. While feedback allows groups to learn about the accuracy of group members, this may be quite difficult in practice. For one thing, the accuracy of members' predictions will depend on the validity of the information on which their prediction is based. For some cue profiles, members may be very accurate, and for some they may not. As such, the weight given to a member's response should be conditional on the information on which that response was based. Estimation of conditional achievement, if attempted in the first place, is more involved than estimating overall achievement. Reliable estimation of conditional achievement requires (many) more observations than a reliable estimation of achievement. It is plausible that by giving feedback, overall achievement, rather than conditional achievement, determines the influence in collective decisions. In this way, the informational value of an individual's response for a given cue profile may be neglected. When there are no large differences in overall achievement, each group member receives the same weight, so that the group process will be identical to a simple majority process. But a simple majority rule neglects the validity of the information on which each response is based. The effect of feedback on group achievement is difficult to predict. While feedback may provide an objective basis to determine an individual's weight in a collective decision, and as such lead to better results than more subjectively determined weights, in practice, important differences in the validity of information for individual decisions may be overlooked, so that objectively determined weights will lead to poorer performance than confidence-based weights, since differences in the validity of information may result in differences in confidence.

Another change in this experiment was that the number of trials in the individual task was increased from 100 to 200. While the individual scores did not increase significantly over the last 50 trials in the first experiment, so that it may be that participants reach their asymptotic performance after 50 trials, increasing the number of trials should give participants the opportunity to learn more about the relations between cues and event. This might result in an increase in performance and the consistency of individual responses. Also, and more importantly, participants were expected to have a stronger idea about their conditional achievement for different partial cue profiles. Hence, the relation between confidence and conditional achievement was expected to be stronger than in the first experiment. A final change compared to the first experiment was in the scale of the confidence ratings. In experiment 1,

a nine-point rating scale was used. While such a scale may give participants the opportunity to provide a reasonably precise indication of their confidence, the influence of confidence in the group process may be different if participants relate confidence more directly to their subjective evaluation of the probability that their response is correct. Therefore, confidence ratings were now to be given on a 100-point scale, with instructions to use it like a subjective probability scale.

Recapitulating, the main purpose of the second experiment was to investigate the group process in the situation of partly shared, partly unique information. The validity of the shared information is not expected to result in differences in group process. In both conditions, groups were expected to indirectly function under a weighting-by-achievement process, and directly under a weighting-by-confidence process. While no difference in group process was expected, the shared information is expected to have a greater impact on the collective decisions when its validity was relatively high. As such, the common knowledge effect is expected in the condition where the validity of the shared information is relatively high, but not in the condition where the validity is relatively low.

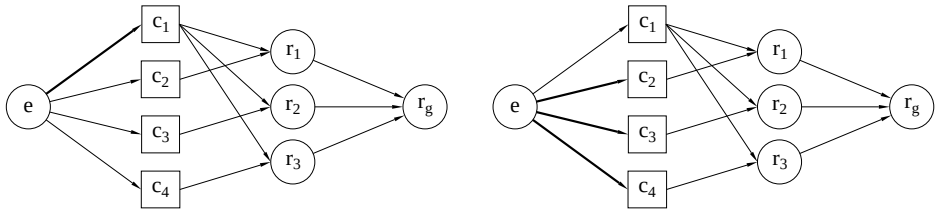
5.4.1 Method

Participants and design

Eighty-seven university undergraduates participated in the experiment. There were 36 males and 51 females. The mean age was 21.92 ($SD = 4.24$). The experiment had a 2 (validity of shared information) $\times$ 2 (feedback in group trials) factorial between-subjects design. The first factor concerned the validity of the shared information, which was either high, when the shared cue had identical properties to c_1 in Table 5.1, or low, when the shared cue had identical properties to c_2 in Table 5.1. The second factor concerned whether feedback on the actual disease was given during the group trials. There were 8 three-person groups ($n = 24$) assigned to the high-validity/no-feedback condition. The other three conditions each had 7 three-person groups ($n = 21$).

Task

The NMCPL task was similar to that in experiment 1. Participants had to decide whether patients had disease A or B on the basis of the presence or absence of a number of symptoms. These symptoms were labelled 'fever', 'headache', 'vomiting', and 'diarrhea'. The total ecological system consisted of four cues and a criterion, but in the individual task, participants were trained in a partial ecological system consisting of two cues (one like c_1 and one like c_2 in Table 5.1) and a criterion. Hence, the partial ecological system of the individual task was identical to that in experiment 1. In the group task, the full ecological system was presented to all participants. The total ecological system was different for the high-validity and low-validity conditions. In the high-validity conditions, there was one cue like c_1 and three cues like c_2 in Table 5.1, while in the low-validity conditions, there was one cue like c_2 and three cues like c_1 in Table 5.1. While the validity of the shared cue was higher in the high-validity conditions, the total ecological system was of less validity. This



(a) High Validity (b) Low Validity

Figure 5.3: Group task systems in experiment 2. Thick arrows between e and c_j indicate relatively high validity ($\eta_j = .28$) and thin arrows relatively low validity ($\eta_j = .13$) of the cues.

is reversed in the low-validity conditions, where the shared cue was of less validity but the total ecological system of higher validity. As in experiment 1, it was assumed that group members would base their individual responses on the partial ecological system they encountered in the individual tasks. A graphical representation of the difference between the conditions is given in Figure 5.3. The probabilities defining the two ecological systems are given in Appendix 5E.

Procedure

The procedure was essentially similar to that of experiment 1. However, the individual task now consisted of 200 trials, which were presented in blocks of 50. After completing each block, participants were instructed to wait until the other group members completed that block before they could proceed to the next block. This was done in order to synchronise the time that participants finished the individual task as much as possible, while allowing each participant to work at his/her own pace. The group task consisted of 32 trials (each of the possible 16 combinations of cue values was presented twice). Each trial began with the presentation, in random order, of the values of the four cues. This was followed by an individual response, and a rating of the confidence in that response. In contrast to experiment 1, confidence ratings were now made on a scale from 0 to 100. Participants were instructed to use the scale in the following way. A score of 0 means one is completely sure the given response is incorrect, and hence, one is completely sure the alternative response is correct. A score of 50 would mean one is just as confident in the given response as in the alternative response. A score of 100 would mean one is completely sure the given response is correct. These instructions were meant to direct participants to use the scale as a (subjective) probability scale. For dichotomous decisions, it is customary to use a scale from 50 to 100, since it is irrational to give a response if the probability that the alternative is the correct one is deemed lower than .50 (participants should then have chosen the other alternative). However, participants were asked to rate their confidence in the group response on the same scale. Since the group may choose an

alternative the participant deems very unlikely, participants could indicate this by rating their confidence in the group response as lower than 50. After rating the confidence in their individual response, participants were instructed to discuss the case in the group in order to arrive at a group response. Each participant was then asked to rate their confidence in the group response. After completing the group task, the experiment was concluded with a computerised exit-interview.

Dependent measures

The dependent measures were the same as in experiment 1. The possible individual scores S_i again had a range of $[.2; .8]$. The ecological environment in the group task was different than in experiment 1. For the ecological environment in high-validity conditions, the range of the possible group scores S_g was $[.15; .85]$. For the ecological environment in the low-validity conditions, this range was $[.10; .90]$. Since in the 32 trials of the group task, each cue profile was presented twice, the trials in the group task were not a proper reflection of the base-rates of the cue profiles in the complete environment. In the actual group task, the range of the possible group scores in the high-validity conditions was $[.21; .79]$ and in the low-validity conditions this range was $[.18; .82]$. Besides the individual and group scores, a final dependent measure was Conf, the participants' reported confidence in their individual responses.

5.4.2 Results

The structure of this section is as follows. First, results for individual performance will be discussed. Then, group achievement will be compared between conditions in order to investigate the effect of the validity of shared information and feedback on group achievement. After this, group scores are compared to the expected group scores from a simple majority and weighting-by-achievement process. This is followed by a test of the hypothesis that the common knowledge effect is present in the high-validity, but not the low-validity conditions. Finally, the group process will be investigated more in depth, by comparing the group responses to predictions from a simple majority and weighting-by-achievement process, and an analysis of the influence of member confidence on the collective response.

Individual performance

The individual task consisted of 4 blocks of 50 trials each. To determine learning effects, a 1-factor (Block) repeated-measures ANOVA was performed over the individual scores. As expected, the effect of Block was significant, $F(3, 258) = 25.47, p < .001$. Pairwise comparisons between consecutive blocks showed that the individual scores in block 1 differed significantly from those in block 2, $F(1, 86) = 25.47$, $p < .001$, and the individual scores in block 2 differed significantly from those in block 3, $F(1, 86) = 6.24$, $p = .014$. There was no significant difference between the individual scores in block 3 and 4.

As in experiment 1, the individual scores from the last 50 trials were used in the remaining analyses. The mean individual score in the last 50 trials was .68 ($SD = .12$). This is identical to the mean individual score in the last 50 trial in experiment 1.

Hence, the additional 100 trials did not appear to result in a significant increase in performance.

Cue utilisation coefficients were computed for the responses in the last 50 trials. The mean of υ_0, the utilisation coefficient of the base rate, was .01 ($SD = .02$). Hence, participants did not show a definite preference for one of the response alternatives. The mean of υ_1, the utilisation coefficient of c_1, was .34 ($SD = .30$). This cue had the highest validity ($\eta_1 = .28$), and in order to give optimal responses, participants could respond to this cue alone. Hence, for consistently optimally responding participants, υ_1 would be 1. There were four participants who consistently gave the optimal response. The mean of υ_2 was .12 ($SD = .13$), which is equal to the validity of c_2 ($\eta_2 = .12$). Finally, the mean of υ_{12}, the utilisation coefficient of the cue interaction, was .01 ($SD = .05$). As in experiment 1, participants appeared to base their responses more on separate cues than on cue profiles. The mean consistency coefficient ξ_i was .47 ($SD = .31$). This is about equal to the mean consistency in experiment 1, so the additional trials in the individual task did not appear to have an effect on response consistency.

As initial differences between conditions may have influenced later results, an ANOVA was performed in order to ascertain such differences. This analysis showed no such differences: $F(3, 85) = .44, p = .725$.

Group achievement

The mean group score was .71 ($SD = .02$) in the high-validity/no-feedback condition, .66 ($SD = .07$) in the high-validity/feedback condition, .67 ($SD = .07$) in the low-validity/no-feedback condition, and .69 ($SD = .05$) in the low-validity/feedback condition. An ANOVA showed no significant main effect for Validity, $F(1, 25) = .84, p = .268$, or Feedback, $F(1, 25) = .32, p = .577$, nor a significant Validity×Feedback interaction effect, $F(1, 25) = 2.55, p = .123$.

As in experiment 1, the expected group scores were computed for two group processes. The assumption of consistency in responses from partial to complete ecological system (see section 5.2.2) was tested using the same procedure as in experiment 1, and was rejected for only one participant (for $\alpha = .005$). Since the assumption had to be rejected only once, the derivation of the expected group response matrices from the individual response matrices was deemed valid.

For a simple majority process, the mean expected group score was .68 ($SD = .02$) in the high-validity/no-feedback condition, .68 ($SD = .07$) in the high-validity/feedback condition, .65 ($SD = .08$) in the low-validity/no-feedback condition, and .70 ($SD = .06$) in the low-validity/feedback condition. There were no significant differences between the conditions for the expected group scores, $F(3, 25) = 1.07$, $p = .379$. The observed group scores were compared to the expected group scores in a 2 (Validity) × 2 (Feedback) × 2 (Source: observed or expected) ANOVA with the last factor treated as a within-groups factor. This last factor had no significant effect, $F(1, 25) = .58$, $p = .455$, which indicates that, overall, group scores did not differ significantly from the expected group scores from a SM process. There was a marginally significant interaction between Feedback and Source, $F(1, 25) = 3.34$, $p = .079$, which indicates that for the no-feedback conditions, the observed group scores were higher than

expected, while for the feedback conditions, observed group scores were lower than expected.

For a weighting-by-achievement process, the expected group scores were .71 (SD = .01) for the high-validity/no-feedback condition, .70 (SD = .03) for the high-validity/ feedback condition, .72 (SD = .02) for the low-validity/no-feedback condition, and .71 (SD = .03) for the low-validity/feedback condition. There were no significant differences between the conditions in these expected group scores, $F(3, 25) = .197$, $p = .897$. The observed group scores were compared to the expected group scores with a similar ANOVA as above. There was a significant effect of the Source factor, $F(1, 25) = 6.23$, $p = .020$, which indicates that, overall, observed group scores were lower than expected from a WA process. There were no significant interactions between Source and the other factors.

Cue utilisation and the common knowledge effect

As for the individual responses, cue utilisation coefficients were computed for the group responses. Since the complete ecological environment consisted of four cues, there were a total of 15 utilisation coefficients (4 for the separate cues and 6 for the first-order, 4 for the second-order, and 1 for the third-order cue interactions). Appendix 5F contains all utilisation coefficients for each group. Not all utilisation coefficients are of interest. In order to determine the common knowledge effect, we are mainly interested in υ_1, the utilisation coefficient of the shared cue. Presence of the common effect should have resulted in υ_1 being much higher than the other utilisation coefficients. As can be seen in Appendix 5F, υ_1 was in general much larger in the high-validity conditions (M = .55, SD = .33) than in the low-validity conditions (M = .16, SD = .20). The mean utilisation in the latter condition was actually quite close to that found in the individual task (υ_2 = .12). Remember that cue utilisation coefficients can be interpreted in terms of 'percentage variation explained'. By dividing cue utilisation coefficients by consistency coefficients, the proportion of explained variation due to a cue, relative to the other cues, is obtained. For $\upsilon_{1,g}/\xi_g$, there was still a large difference between the high-validity condition (M = .63, SD = .31) and the low-validity conditions (M = .20, SD = .23). These differences show that, as expected, reliance on the shared cue was larger when it had a high validity than when it had a low validity.

The results above clearly indicate the common knowledge effect is mediated by the validity of shared information relative to unique information. Insofar as the common knowledge effect is caused by anything other than the validity of information, one might expect the shared information to entirely dominate the collective response in the high-validity condition. In this case, the impact of the shared information on the collective decision would be higher than expected from its validity. To test this domination of shared information, two logistic regression models were compared. The first model only had c_1, the shared cue, as a predictor, while the second model had all four cues as predictors. Since the first model is nested in the second, a χ^2-difference test can be used to infer whether the second results in a significantly better fit than the first. If group responses are only dependent on the shared cue, the χ^2-difference test should be non-significant. Overall, the second model did result in a significantly better

Table 5.4: Types of group decisions

	Low validity		High validity	
	No feedback	Feedback	No feedback	Feedback
Minority	22	19	11	11
Majority	126	144	65	114
Unanimous	76	61	180	99

fit, $\chi(3) = 30.33$, $p < .001$. This also held for the low-validity conditions separately, $\chi(3) = 112.32$, $p < .001$, as well as for the high-validity conditions, $\chi(3) = 26.59$, $p < .001$. Hence, there is no strong evidence indicating groups neglected unique information, even when its validity was relatively low[2].

Group process

The frequencies of minority, majority, and unanimous decisions are given in Table 5.4. As can be seen in this table, the number of unanimous decisions was higher in the high-validity than in the low-validity conditions. This was expected, since, due to its larger validity, the shared cue should dominate the individual responses in the high-validity condition. Also, the number of minority decisions was higher in the low-validity than in the high-validity conditions. This could indicate more groups in the former condition functioned under a weighted majority process. As in experiment 1, group responses were compared to predictions following from a simple majority process (SM) and a weighting-by-achievement (WA) process. Groups were classified as either SM, WA, or neither, according to the lowest RMSEP value. In the high-validity/no-feedback condition, 1 group was classified as SM, and 7 as WA. In the high-validity/feedback condition, 6 groups were classified as SM, and 1 as WA. In the low-validity/no-feedback condition, 2 groups could not be classified, 2 were classified as SM and 3 as WA. In the low-validity/feedback condition, 3 groups were classified as SM and 4 as WA. There appears to be no overall effect of Validity on the group process. Both in the high and low-validity conditions, about half of the groups were classified as SM, and half as WA. Collapsing over Validity and ignoring the groups that could not be classified, 3 groups in the no-feedback conditions were classified as SM, and 10 as WA, while in the feedback conditions, 9 groups were classified as SM, and 5 as WA. This is a marginally significant difference, $\chi(1) = 3.12$, $p = .077$. While there are too little groups to provide reliable evidence, the finding that feedback appeared to affect the group process is interesting and requires further research.

Group process and confidence

As before, confidence was expected to play a significant role in the actual group process. The rank-correlation between Conf and $(A_i|C_m^i)$ was $\tau = .28$ ($Z = 21.85$,

[2]Besides these two models, a third model was also fitted, which included the four predictors as well as all their interactions. In neither condition did this model result in a significantly better fit.

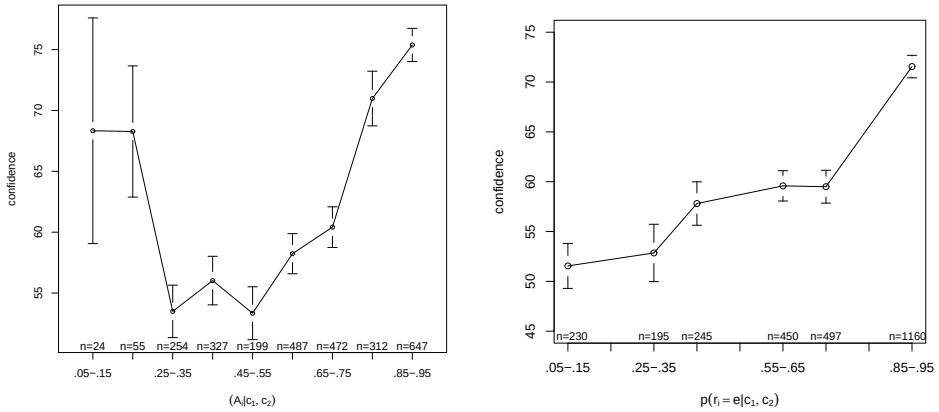


Figure 5.4: Relation between confidence ratings and $(A_i|c_1, c_2)$ and $P(r = e|c_1, c_2)$

$p < .001$). In contrast to experiment 1, the correlation between Conf and $P(e_k|C_m^i)$ was just as high, $\tau = .28$, ($Z = 22.10$, $p < .001$). The relation between the confidence ratings and both $(A_i|C_m^i)$ and $P(e_k|C_m^i)$ respectively is depicted in Figure 5.4.

The means and standard deviations of Conf for minority and majority group members, for those situations in which there was initial disagreement ($n = 512$), are given in Table 5.5. A 2 (Validity) × 2 (Feedback) × 2 (Decision Type) × 2 (Source) ANOVA with WLS estimation was performed for this data. This analysis showed no significant main effects, but as expected, a significant interaction between Decision Type and Source was found, $F(1, 1518) = 64.65$, $p < .001$. Two post-hoc t-tests (with a Welch-correction for the degrees of freedom) showed that minority confidence was significantly higher than majority confidence for minority decisions, $t(118.31) = 6.10$, $p < .001$, while minority confidence was significantly lower than majority confidence for majority decisions, $t(970.32) = -7.17$, $p < .001$. Besides this effect, the ANOVA showed a significant Validity × Source interaction, $F(1, 1518) = 12.60$, $p < .001$, and a Decision Type × Feedback × Source interaction, $F(1, 1518) = 7.97$, $p = .005$. This was an unexpected result, but it indicates that the difference between minority and majority confidence was larger when no feedback was given. As in experiment 1, the results are consistent with a weighted majority process. When group decisions followed the minority position, the confidence of the minority was much higher than the confidence of the majority. When group decisions followed the majority position, the confidence of the minority was lower than that of the majority. The effect of confidence appears larger when no objective feedback was given. This is consistent with the idea that feedback is used as a means to determine member competence. Hence, in the presence of feedback, there was less need to rely on member confidence in the group process.

In order to ascertain how confidence directly influenced group decisions, the three logistic regression models also used in experiment 1 were compared. The first model, representing a simple majority group process, was identical to (5.10). The second model, representing a weighted majority group process, was identical to the model

Table 5.5: Confidence ratings of minorities and majorities for minority and majority decisions

	Low Validity				High Validity			
	No Feedback		Feedback		No Feedback		Feedback	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
				Minority decisions				
Minority	75.32	16.43	71.00	21.76	66.82	13.09	55.55	16.50
Majority	49.64	16.10	52.74	16.80	49.09	18.81	57.27	20.74
				Majority decisions				
Minority	54.94	20.14	59.07	19.88	50.80	14.13	55.25	18.47
Majority	60.21	22.88	64.94	20.23	67.79	17.38	64.37	20.24

given in (5.11), but now with $\max(x) = 100$. The last model represented the weighted majority process with weights being an exponential function of confidence, and was identical to that given in (5.11), but now with $\max(x) = 100$ and $\overline{x} = 63.27$. Table 5.6 contains the results of these three models. As can be seen, M_2 provided the best fit and was selected over the other two models. Hence, the group process appears to have followed a weighted majority rule, with weights being a linear function of confidence. As in experiment 1, the analysis was repeated for each condition separately, again with 2-parameter models with a single predictor consisting of the sum of the original predictors. The results of these analyses are given in the lower part of Table 5.6. The model comparison does not show an entirely clear pattern. The selected models were M_2 for the high-validity/no-feedback condition, M_1 for the high-validity/feedback condition, M_2 for the low-validity/no-feedback condition, and M_3 for the low-validity/feedback condition. Combined with the results from the comparison of predicted and observed group responses reported in section 5.4.2, there is clear evidence that groups in the high-validity/feedback condition functioned under a simple majority process. In the other conditions, groups did appear to weight members' contributions according to confidence. As such, there is also more indirect evidence for a weighting-by-achievement process in these conditions.

Discussion

As expected, there was evidence for the common knowledge effect in the high-validity conditions, but not in the low-validity conditions. As such, the common knowledge effect appeared to be mediated by the validity of shared information, relative to that of unique information. Further support for the view that the effect should be attributed to the validity of information, comes from the finding that groups in the high-validity conditions did not ignore unique information, since this information was also valid, although to a lesser extent than the shared information. The validity of shared information, or the presence of feedback, did not appear to have an effect on the group achievement. Inspection of the means does appear to indicate the presence of an interaction, with feedback resulting in an increase of the group score for groups

Table 5.6: Logistic regression models for $P(g = 1)$

	M_1			M_2			M_3		
	est.	Z	p	est.	Z	p	est.	Z	p
α	-0.06	0.14	0.67	-0.12	-0.83	0.41	-0.11	-0.74	0.46
β_1	2.01	12.63	$< .001$	3.75	11.53	$< .001$	2.25	11.99	$< .001$
β_2	2.00	12.50	$< .001$	3.43	11.53	$< .001$	2.12	12.00	$< .001$
β_3	2.04	12.47	$< .001$	3.93	11.26	$< .001$	2.32	11.80	$< .001$
$-2 \log L$	382.97			322.80			327.08		
				High-validity/no-feedback					
α	-0.23	-0.67	.50	-0.28	-0.65	.52	-0.29	-0.70	.48
β	1.96	6.78	$< .001$	3.62	5.74	$< .001$	2.12	6.19	$< .001$
$-2 \log L$	63.14			42.13			48.54		
				High-validity/feedback					
α	-0.07	-0.21	.84	-0.03	-0.10	.93	-0.06	-0.20	.84
β	2.36	7.64	$< .001$	3.32	6.92	$< .001$	2.21	7.15	$< .001$
$-2 \log L$	74.42			81.48			75.50		
				Low-validity/no-feedback					
α	0.00	0.00	.99	-0.09	-0.33	.75	-0.06	-0.23	.82
β	1.80	8.23	$< .001$	3.96	6.85	$< .001$	2.24	7.58	$< .001$
$-2 \log L$	125.17			99.06			102.21		
				Low-validity/feedback					
α	-0.02	-0.06	.95	-0.17	-0.67	.50	-0.12	-0.47	.64
β	2.05	8.60	$< .001$	3.71	7.28	$< .001$	2.28	7.87	$< .001$
$-2 \log L$	117.63			102.34			101.73		

in the low-validity condition, and a decrease in the high-validity condition. The non-significance of the interaction effect may have been due to insufficient power. Further research will be needed to show whether this supposed interaction is actually present. The presence of such an interaction is plausible when considering that the presence of feedback may affect the social process by which a group decides on a collective response. This is consistent with the analyses of the group process: without feedback, groups mostly worked under a weighting-by-achievement rule, while with feedback, groups mostly worked under a simple majority rule. This effect of feedback was mostly noticeable in the high validity conditions. This is also apparent when looking at the influence of confidence in the group process. In the high-validity/feedback condition, groups did not appear to weight members' contributions by their confidence while in the other conditions, there was evidence for such a weighting process. That feedback results in a simple majority process is consistent with a possible effect of feedback proposed in section 5.4. Here it was argued feedback might lead groups to neglect differences in the validity of information between cue profiles. If the overall achievement of individual group members is similar, a weighting by overall achievement will be identical to a simple majority process. If this is so, then the finding that groups in the low-validity/feedback condition did appear to weight by member confidence should be explained by a higher variance in overall achievement in this condition. However, such a difference was not found. The difference in the effect of feedback on the group process between the high and low-validity conditions may be explained otherwise. Since the validity of the shared cue in the latter condition is lower than that of the unique cues, and there is more initial disagreement, differences in the

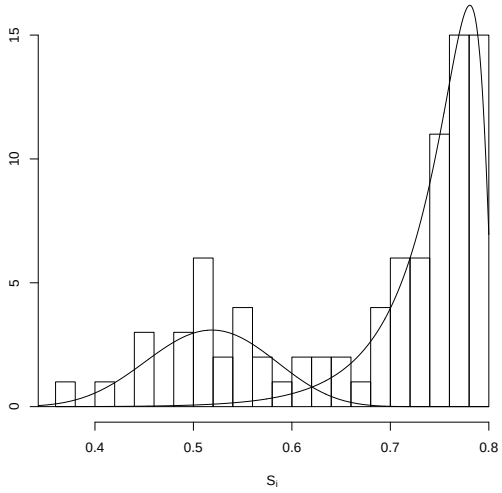


Figure 5.5: Histogram and fitted mixture distribution for S_i in the last 50 trials. Values on the y-axis are frequencies, values on the x-axis are individual scores.

validity of partial cue profiles may be more apparent in this condition than in the high-validity condition. As such, groups may be more persuaded to consider such differences in validity when forming a group response.

Surprisingly, increasing the number of trials in the individual task from 100 to 200 did not result in a higher mean of the individual scores in the last trials as compared to the first experiment, nor in an increase in the consistency of individual responses. However, inspection of the distribution of individual scores in the present experiment indicates a mixture distribution. This was confirmed in a mixture analysis for the logit-transformed individual scores[3]. The best fitting mixture was a 2-component Normal-mixture with unequal variances. Figure 5.5 contains a histogram of the individual scores, with the fitted mixture distribution (transformed back to the original scale) superimposed. On the original scale, the estimated mean individual score in the first component was .76 and in the second it was .54. The mixture is easily interpreted as consisting of a group which mastered the environment ($n = 62$) and a group which did not ($n = 25$). The number of individuals classified as non-learners is relatively high, which may be due to a lack of motivation or the inherent difficulty of learning in the environment. That the mean consistency of individual responses was also about equal to that found in the first experiment can be attributed to the same cause as that underlying the equivalence of the individual scores. Taking the classification as 'learners' and 'non-learners' from the mixture analysis into account, the mean consistency coefficient for the learners was .62 ($SD = .21$), while the mean consistency of the non-learners was .09 ($SD = .10$). As such, the additional trials did appear to have an effect on the consistency of a large number of participants. There

[3]The mixture analysis was performed with MCLUST (Fraley & Raftery, 2002). In the analysis, 10 models were compared. These were 1 to 5 component normal-mixtures with equal or unequal variances. Model selection was based on C_{BIC} (see Chapter 3).

Table 5.7: Mean group and individual scores of experiment 1

	Distributed		Shared	
	M	*SD*	*M*	*SD*
Group score	.62	.06	.58	.05
Worst member	.51	.06	.52	.05
Average member	.57	.04	.57	.03
Best member	.63	.04	.61	.03

was no clear indication of a mixture distribution in the first experiment, although this does not warrant the conclusion that there was no separation between 'learners' and 'non-learners' in this experiment also. Increasing the number of trials at least results in a clearer separation of 'learners' and 'non-learners'.

5.5 Groups vs individuals

In light of the general theme of this thesis, it is interesting to determine whether groups outperform individuals when it comes to making good decisions. Therefore, group achievement is compared to the performance of individual group members in this section. There are at least three possible comparisons, those between group achievement and performance of the worst, average, and best group members respectively. For each group in the two experiments, the individual performance of the worst, average and best group member were determined. For experiment 1, the means and standard deviations of these three individual scores and the group score are given in Table 5.7. As can be seen in this table, on average, group achievement lies between the performance of the average and best group member. The group scores were compared to the best individual scores in a 2 (Condition) $\times$ 2 (Source: group vs individual) ANOVA, with the last factor as a within-groups factor. This analysis showed no significant effect of Source, nor a Source $\times$ Condition interaction. Hence, it should be concluded that the group scores were not significantly different from the individual scores of the best group member in both conditions. In contrast, when groups were compared to the average individual, there was a significant effect of Source, $F(1, 28) = 28.54$, $p < .001$, as well as a significant Source$\times$Condition interaction, $F(1, 28) = 7.24$, $p = .01$. Groups were advantaged over the average group member, and moreover, this advantage was larger in the distributed than in the shared condition.

The individual and group scores for experiment 2 are given in Table 5.8. Group scores were compared to the individual scores of the best group members by a 2 (Validity) $\times$ 2 (Feedback) $\times$ 2 (Source: group vs individual) ANOVA, with the last factor as a within-group factor. This analysis showed no significant effects. Hence, group scores were not significantly different to the individual scores of the best performing group members. In contrast, when comparing group scores to the average individual score of group members, the ANOVA showed a significant main effect of

Table 5.8: Mean group and individual scores of experiment 2

	Low-validity				High-validity			
	No-feedback		Feedback		No-feedback		Feedback	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group score	.67	.07	.69	.05	.71	.02	.66	.07
Worst member	.53	.10	.54	.05	.63	.06	.54	.04
Average member	.61	.04	.62	.03	.67	.03	.62	.05
Best member	.68	.01	.67	.03	.71	.02	.69	.06

Source, $F(1, 25) = 30.70$, $p < .001$, from which it should be concluded that group scores were significantly higher than the average individual scores of the group members. Also, there was a significant Source × Validity interaction, $F(1, 25) = 11.74$, $p < .01$, a significant Source × Feedback interaction, $F(1, 25) = 4.29$, $p = .05$, and a significant Source × Validity × Feedback interaction, $F(1, 25) = 4.89$, $p = .04$. The first interaction indicates that the advantage of groups over average individuals was larger in the high than in the low-validity conditions. The second interaction indicates that the advantage of groups over average individuals was larger in the no-feedback than in the feedback conditions. Finally, the three-way interaction indicates that the largest advantage of groups over the average group member is found in the high-validity/no-feedback condition.

5.6 General discussion

Two experiments investigated collaboration in Nonmetric Multiple Cue Probability Learning (NMCPL) under different distributions of information over group members. In the first experiment, information was either completely distributed or completely shared. As expected, group achievement was higher in the first as compared to the latter condition. In the second experiment, information was partly shared, partly unique. Here, the relative validity of the shared and unique information did not appear to affect group achievement.

As Steiner (1972) noted, group achievement depends on task requirements, group resources, and group process. NMCPL tasks are optimising tasks. The objective is to optimally predict an event on the basis of cues which are probabilistically related to the event. If the cues are conditionally independent, as they were in the ecological systems of the experiments, NMCPL tasks are divisible tasks. This means a division of labour can be profitable. An important reason to divide cognitive labour is that there are limits to an individual's capacity to process information. In realistic situations, there are many cues which are possible indicators of an event. Learning about the relations between cues and event will be impeded when individuals consider all possible cues as relevant. By restricting the information that is processed, by focussing on a subset of cues and ignoring others, learning will be facilitated. This will result in individuals responding more optimally to their information. In principle,

groups are able to perform optimally when the resources are completely distributed over group members. Hence, a division of labour makes it possible for a group to increase its informational base, while staying within the limits of individual capacity. A division of labour does put rather stringent requirements on the group process. When resources are completely distributed, as in the distributed condition in experiment 1, the optimal group process is a weighting-by-evidence process, in which each individual response is weighted by its evidential support. This process requires precise knowledge of the structure of the (partial) ecological system. If individuals had such precise knowledge, one would expect them to always respond optimally to their information. This consistently optimal response behaviour is not often encountered. While precise knowledge of the structure of the (partial) ecological system is likely to be missing, individuals may have knowledge regarding their ability to give correct predictions to specific information. This knowledge can be used in a weighting-by-achievement process, in which each individual response is weighted according to individuals' track-records of giving correct predictions to specific information. When the group process is a weighting-by-achievement process, groups will suffer some process loss. However, this process loss is less than the loss when the group process is a simple majority process. As its name indicates, this latter process is simple to apply. If information is shared, so that all group members base their responses on identical information, each individual has the same potential of making a correct prediction. In such a situation, a simple majority process will often be adequate to realise potential group achievement. However, if group size is relatively small and there are large differences in individuals' predictive accuracy, a weighting-by-achievement process will outperform a simple majority process. In the case that information is partly shared and partly distributed, as it was in the second experiment, a weighting-by-achievement process may actually outperform a weighting-by-evidence process. This is because the weighting-by-evidence process will put too much weight on the shared cue. Overall, weighting-by-achievement is a viable and reasonably adequate group process. It also seems more plausible than the weighting-by-evidence process.

As indicated earlier, the weighting-by-achievement process defined in (5.7) should not be taken as a model of the actual way in which groups arrive at a collective decision. It should be viewed as a somewhat idealised representation of group members' influence in the collective decision. In practice, this influence will be more directly determined by members' confidence in their individual decisions. Confidence was assumed to be dependent on achievement, and this assumption was confirmed in both experiments. Overall, the group process was adequately described by a model in which the individual decisions are weighted by confidence. Due to the relation between confidence and achievement, it was thought the group process would resemble a weighting-by-achievement process more than a simple majority process. In the first experiment however, there were more groups in which the group process resembled a simple majority process than groups in which the process resembled a weighting-by-achievement process. In the second experiment, the group process resembled a simple majority process for about half the groups, and a weighting-by-achievement process for the other half. It should be noted that the predictions on which these classifications are based are derived entirely from group members' response behaviour in the individual task. Hence, these predictions are accurate if individuals' response

behaviour is consistent over the individual and group task. In the first experiment, the individual task consisted of less trials than the individual task in the second experiment. For this reason, participants in the first experiment are expected to be less consistent than participants in the second experiment. While this is not immediately clear from the mean consistency coefficients in both experiments, there was a clear indication of a mixture distribution in the second experiment. Consistency coefficients for 'learners' in the second experiment were clearly higher than the mean consistency score in the first experiment, where there was no clear indication of a mixture distribution. Also, the assumption of consistency over individual and group task was more often rejected in the first experiment than in the second. As such, the predictions from a weighting-by-achievement process are more likely to be valid in the second experiment than in the first. Predictions from a simple majority process are less influenced by inconsistent response behaviour, because the predictions from the weighting-by-achievement process are generally more extreme (closer to either 1 or 0). Therefore, the finding that not many groups in the first experiment could be classified as functioning under a weighting-by-achievement process may be attributable to the relative inconsistency of individual response behaviour in this experiment.

The results regarding the group process were somewhat mixed. Overall, there was clear evidence that the members' influence in the collective decisions depended on their confidence, and that confidence was related to achievement. But for a large number of groups the underlying group process was more alike to a simple majority process than a weighting-by-achievement process. This divergence between inference of the group process based on predictions from individual response behaviour and inference based on member confidence also arises when looking at the relation between the distribution of information and group process. When information was completely shared, as in the shared condition in the first experiment, there was no evidence that members' contributions were weighted by their confidence. When information was (partly) distributed, groups did weight members' contributions by their confidence. When inferring the group process on the basis of predictions from individual response behaviour, the relation between information distribution and group process was less marked. There was no clear difference in the underlying group process between the situation of completely distributed and that of completely shared information. In the situation of partly shared, partly unique information, there were more groups classified as weighting-by-achievement. As mentioned earlier, the classification in this experiment may be more valid, due to the higher consistency in individual response behaviour. There was some indication that the presence of feedback had an influence on the group process. This was mainly visible in the high-validity conditions, where almost all groups were classified as weighting-by-achievement when no feedback was given, and almost all groups were classified as simple majority when feedback was given. Relating this to the obtained group scores, the feedback appeared to have a detrimental effect on group achievement, since weighting-by-achievement is the more optimal group process. The reason why feedback moved groups to adopt a simple majority process is not directly evident.

Another reason why inferences based on model predictions and inferences based on confidence diverge lies in the imperfect relation between confidence and achievement. If confidence was perfectly related to achievement, weighting-by-achievement

would be similar to weighting-by-confidence, but otherwise the two will lead to different results. As noted before, this leads to two causes of process loss. The first is the difference between the underlying group process (e.g. weighting-by-achievement) and the optimal group process (e.g. weighting-by-evidence), the second is the difference between the underlying group process (e.g. weighting-by-achievement) and the actual group process (e.g. weighting-by-confidence). With the methods used in these experiments, the two causes can be clearly separated. Process loss due to the first cause can be determined by comparing the predicted group scores from different group processes to the maximum group score, while process loss due to the second cause can be determined by comparing the obtained group scores to the predicted group scores. Overall, the bulk of process loss is attributable to the first cause. Actual group scores were much closer to the predicted group scores than predicted group scores to the maximum group scores.

The main purpose of the two experiments was to determine whether groups outperform individuals in NMCPL tasks, and how the advantage of groups over individuals is determined by the two causes mentioned in the beginning of the chapter. In the first experiment, conditions were created so that these two causes could be separately investigated. The advantage of groups over individuals due to the greater informational base of groups could be determined in the distributed condition. The advantage of groups over individuals due to minimising the effect of idiosyncratic bias could be determined in the shared condition. Since group achievement in the distributed condition was higher than that in the shared condition, the effect of the first cause is larger than that of the second. In the second experiment, information was partly shared and partly unique, and the two causes could not be investigated separately. A possible concern in this situation is the common knowledge effect, which leads groups to be more influenced by shared than by unique information and would eliminate the advantage due to a larger informational base. A plausible explanation of this effect can be given in terms of social validation. If the validity of information for the problem at hand is unknown, the number of people sharing the information may be an indicator of the validity of the information. If the common knowledge effect is caused by social validation, it is not expected when group members have prior information on the validity of information. The results of the second experiment show the weight given to shared information is indeed related to the validity of the information. The influence of the shared cue was much higher when it had a high validity than when it had a low validity, relative to the validity of the unique information. However, this did not result in groups neglecting unique information in the favour of shared information. As such, the common knowledge effect may not be of much concern when group members have a good idea of the validity of information.

In both experiments, group performance was about equal to the performance of the best group member. While this does not show groups are advantaged over all group members, it does show groups are advantaged over the average group member. If information is completely distributed, potential group performance is higher than the potential performance of any group member, no matter how optimal their responses are for their partial cue profiles. Since group achievement was not better than the achievement of the best group member, groups were unable to realise this potential fully. If this result holds in general, the question of whether groups outperform

individuals cannot be answered unequivocally. The answer is 'yes' when reference is to the average individual, but 'no' when reference is to the best performing individual. While there is no apparent loss associated to letting groups make a decision rather than an individual, there is no apparent gain either if it is possible to pick the best performing individual beforehand. When it comes to decision-making, the effects of the 'cancelling-out' of idiosyncratic bias and the larger informational base of groups as compared to individuals are not large enough to raise group performance above that of all group members. While this is a negative finding, the result groups perform at the level of the best individual group member shows such relatively high group performance is not reserved for tasks which meet the demonstrability requirements, as is often thought.

In addition to the empirical findings, this chapter has offered novel methods for studying individual and collective behaviour in NMCPL. The information-theoretic cue utilisation coefficients were useful for the analysis of individual and group cue use. A full-blown information-theoretic analysis of individual and collective behaviour in NMCPL is promising and certainly possible, but hypothesis testing in this framework will require further investigation of the distribution of the information-theoretic measures (more specifically, on the distribution of the interaction-measures). Compared to previous research on interpersonal learning in MCPL tasks, the experimental design applied here has large advantages. By changing the ecological system from individual to group task, previous research on interpersonal learning actually manipulated individual bias. As such, adequate group performance required participants to unlearn what they previously learned, rather than learning additional aspects of the ecological system from each other. By training individuals in partial ecological systems, a more realistic situation is created where individuals may learn from each other about additional aspects of the ecological system. There are many possible variations in the particulars of the ecological system and the distribution of information over group members. Hopefully, the paradigm employed here will open the door for many interesting studies on interpersonal learning in realistic situations.

Appendix

5A Cue validity and utilisation coefficients

This appendix derives cue validity and cue utilisation coefficients from classical measures of information (Shannon, 1948). First, a short overview of the properties of entropy and information measures is given, after which the coefficients are derived.

Suppose an event e_k, $k = 1, \ldots, K$ has to be predicted. Without any knowledge, the occurrence of each event should be deemed equally likely. That is, the probability $P(e_k) = \frac{1}{K}$ for all k. This null-model has a corresponding entropy of

$$H_0(e) = \log K.$$

In general, the events e_k are not all equally likely, and the entropy is defined as

$$H(e) = -\sum_{k=1}^{K} P(e_k) \log P(e_k).$$

The informational value of knowing the base-rates of events e_k is taken as a reduction in uncertainty or entropy

$$I(e) = H_0(e) - H(e).$$

For two variables e_k and c_j, the joint entropy is defined as

$$H(c, e) = -\sum_{j=1}^{J} \sum_{k=1}^{K} P(c_j, e_k) \log P(c_j, e_k),$$

for which the following relation holds

$$0 \leq H(c, e) \leq H(c) + H(e),$$

with equality on the right-hand side only if c and e are statistically independent. Besides the joint entropy, it is also possible to define the conditional entropy as

$$H(e|c) = -\sum_{j=1}^{J} \sum_{k=1}^{K} P(c_j) P(e_k|c_j) \log P(e_k|c_j).$$

This is equivalent to

$$H(e|c) = H(c, e) - H(c),$$

from which it follows that $H(e|c) = H(e)$ if e and c are independent. The mutual information between e and c is defined as

$$I(c; e) = H(e) - H(e|c) = H(c) - H(c|e) = H(c) + H(e) - H(c, e),$$

i.e. the reduction in uncertainty resulting from knowledge about c. Classical information theory has been formulated for the two variable case (representing one transmitter and one receiver), but McGill (1954) has given a useful extension to a multivariate case (with multiple transmitters and one receiver). In general, the entropy of e can be partitioned as

$$H(e) = I(c_1, \ldots, c_J; e) + H(e|c_1, \ldots, c_J),$$

i.e. as a component of information transmitted by the cues and a component which is unexplained by the cues. As McGill shows, it is possible to partition the 'explained component' further. For instance, for a three variable case (two cues and one event):

$$I(c_1, c_2; e) = I(c_1; e) + I(c_2; e) + A(c_1, c_2, e), \tag{5A.1}$$

in which

$$A(c_1, c_2, e) = I(c_1; e|c_2) - I(c_1; e) = I(c_2; e|c_1) - I(c_2; e).$$

The conditional mutual information $I(c_1; e|c_2)$ is defined as

$$I(c_1; e|c_2) = H(e|c_2) - H(e|c_1, c_2) = H(e|c_1) - H(e|c_1, c_2).$$

Equation 5A.1 can be interpreted in a similar way as an ANOVA-model, with the first two terms representing main effects and the third an interaction effect. The method is easily generalised to situations with more predictors. For three cues, the partitioning of the total transmitted information would be

$$\begin{aligned} I(c_1, c_2, c_3; e) = {} & I(c_1; e) + I(c_2; e) + I(c_3; e) + A(c_1, c_2, e) + A(c_1, c_3, e) + A(c_2, c_3, e) \\ & + A(c_1, c_2, c_3, e), \end{aligned}$$

i.e. into all main effects, all two-way interactions and the three-way-interaction.

The definition of cue validity proposed here is in terms of relative uncertainty reduction. As indicated at the start of this appendix, the maximum uncertainty associated with a dichotomous variable is $H_0 = \log 2$. Cue validity coefficients can be computed by dividing the different $I(\cdot)$ and $A(\cdot)$ components by $H_0(e)$. For instance, the validity of the base rate is defined as

$$\eta_0 = \frac{I(e)}{H_0(e)}, \tag{5A.2}$$

the validity of cue 1 as

$$\eta_1 = \frac{I(c_1; e)}{H_0(e)}, \tag{5A.3}$$

and the validity of the cue interaction between c_1 and c_2 as

$$\eta_{12} = \frac{A(c_1, c_2, e)}{H_0(e)}. \tag{5A.4}$$

Defined in this way, all validity coefficients lie in the interval [0;1], and $\sum_i \eta_i \leq 1$. Furthermore, note that $\sum_i \eta_i = \frac{I(c_1, \ldots, c_J; e)}{H_0(e)}$. As such, the proportion of uncertainty reduction (or 'explained variation') in the entire ecological system is equal to the sum of the validity coefficients. This leads to the definition of an overall predictability coefficient as

$$\psi = \sum_j \eta_j. \tag{5A.5}$$

Cue utilisation coefficients υ_j are defined in an identical way to the validity coefficients η_j, but now the criterion is not the event e_k, but the response r_i. Similar to the predictability coefficient ψ, an individual consistency coefficient ξ_i is defined as the sum of all utilisation coefficients υ_j for individual i.

5B Optimal group process

In a dichotomous choice situation with events e and responses r

$$e, r \in \{-1, 1\}$$

and a group of n independently responding individuals each with a probability $p_i = P(r = e)$ that their response is correct, Nitzan and Paroush (1982) have proved that a weighted majority rule

$$r_g = \operatorname{sgn}\left(\sum_{i=1}^{n} w_i r_i\right),$$

with weights $w_i \propto \log\left(\frac{p_i}{1-p_i}\right)$ is optimal, in the sense that it maximises the probability $P(r_g = e)$. While they do not mention this, their result is intuitively plausible, since this particular weighted majority rule is equivalent to a likelihood-ratio test, which, from the Neyman-Pearson Lemma, are well known to be optimal for choosing between two simple hypotheses. This correspondence can be shown by taking the log-odds transform of p_i

$$\omega_i = \log\left(\frac{p_i}{1-p_i}\right).$$

The log-likelihood ratio of alternative e_1 over e_{-1} is then given as

$$\log\frac{p(r_1,\ldots,r_n|e_1)}{p(r_1,\ldots,r_n|e_{-1})} = \log\left[\prod_i \left(\frac{p_i}{1-p_i}\right)^{r_i}\right] = \sum_{i=1}^{n} r_i\omega_i.$$

Hence, the weighted majority rule is equivalent to a decision on the basis of a likelihood-ratio test when the alternatives have identical prior probabilities and the costs of type I and II errors are equal.

5C Complete ecological system in experiment 1

Table 5C contains the details of the complete ecological system used in experiment 1. The cues c_j are conditionally independent given the events e_k. For each cue profile $C_m = (c_{1:x},\ldots,c_{6:z})$, the corresponding probabilities $P(e_1|C_m)$ and $P(e_2|C_m)$ are given. Also, the probabilities $P(e_1|C_m^i)$ for the partial cue profiles $C_m^1 = (c_{1:x}, c_{2:y})$, $C_m^2 = (c_{3:x}, c_{4:y})$ and $C_m^3 = (c_{5:x}, c_{6:y})$ are given. Note that the values of $P(e_2|C_m^i)$ are not given, but they can easily be reconstructed from the information provided, since $P(e_2|C_m^i) = 1 - P(e_1|C_m^i)$. An asterix ('*') in front of a row number indicates that the particular cue profile was used in the group trials.

Table 5C: Ecological system of experiment 1

m	c_1	c_2	c_3	c_4	c_5	c_6	$P(e_1\|C_m)$	$P(e_2\|C_m)$	$P(e_1\|C_m^1)$	$P(e_1\|C_m^2)$	$P(e_1\|C_m^3)$
*1	0	0	0	0	0	0	.11	.89	.33	.33	.33
*2	1	0	0	0	0	0	.67	.33	.89	.33	.33
3	0	1	0	0	0	0	.02	.98	.08	.33	.33
4	1	1	0	0	0	0	.25	.75	.57	.33	.33
*5	0	0	1	0	0	0	.67	.33	.33	.89	.33
*6	1	0	1	0	0	0	.97	.03	.89	.89	.33
*7	0	1	1	0	0	0	.25	.75	.08	.89	.33
8	1	1	1	0	0	0	.84	.16	.57	.89	.33
9	0	0	0	1	0	0	.02	.98	.33	.08	.33
10	1	0	0	1	0	0	.25	.75	.89	.08	.33

continued on next page

Ecological system of experiment 1 (continued)

m	c_1	c_2	c_3	c_4	c_5	c_6	$P(e_1\|C_m)$	$P(e_2\|C_m)$	$P(e_1\|C_m^1)$	$P(e_1\|C_m^2)$	$P(e_1\|C_m^3)$
11	0	1	0	1	0	0	.00	1.00	.08	.08	.33
12	1	1	0	1	0	0	.05	.95	.57	.08	.33
13	0	0	1	1	0	0	.25	.75	.33	.57	.33
14	1	0	1	1	0	0	.84	.16	.89	.57	.33
15	0	1	1	1	0	0	.05	.95	.08	.57	.33
*16	1	1	1	1	0	0	.47	.53	.57	.57	.33
*17	0	0	0	0	1	0	.67	.33	.33	.33	.89
18	1	0	0	0	1	0	.97	.03	.89	.33	.89
19	0	1	0	0	1	0	.25	.75	.08	.33	.89
*20	1	1	0	0	1	0	.84	.16	.57	.33	.89
21	0	0	1	0	1	0	.97	.03	.33	.89	.89
*22	1	0	1	0	1	0	1.00	.00	.89	.89	.89
23	0	1	1	0	1	0	.84	.16	.08	.89	.89
24	1	1	1	0	1	0	.99	.01	.57	.89	.89
*25	0	0	0	1	1	0	.25	.75	.33	.08	.89
26	1	0	0	1	1	0	.84	.16	.89	.08	.89
27	0	1	0	1	1	0	.05	.95	.08	.08	.89
28	1	1	0	1	1	0	.47	.53	.57	.08	.89
29	0	0	1	1	1	0	.84	.16	.33	.57	.89
30	1	0	1	1	1	0	.99	.01	.89	.57	.89
*31	0	1	1	1	1	0	.47	.53	.08	.57	.89
32	1	1	1	1	1	0	.93	.07	.57	.57	.89
33	0	0	0	0	0	1	.02	.98	.33	.33	.08
34	1	0	0	0	0	1	.25	.75	.89	.33	.08
35	0	1	0	0	0	1	.00	1.00	.08	.33	.08
36	1	1	0	0	0	1	.05	.95	.57	.33	.08
37	0	0	1	0	0	1	.25	.75	.33	.89	.08
*38	1	0	1	0	0	1	.84	.16	.89	.89	.08
39	0	1	1	0	0	1	.05	.95	.08	.89	.08
*40	1	1	1	0	0	1	.47	.53	.57	.89	.08
41	0	0	0	1	0	1	.00	1.00	.33	.08	.08
42	1	0	0	1	0	1	.05	.95	.89	.08	.08
*43	0	1	0	1	0	1	.00	1.00	.08	.08	.08
44	1	1	0	1	0	1	.01	.99	.57	.08	.08
45	0	0	1	1	0	1	.05	.95	.33	.57	.08
46	1	0	1	1	0	1	.47	.53	.89	.57	.08
47	0	1	1	1	0	1	.01	.99	.08	.57	.08
48	1	1	1	1	0	1	.13	.87	.57	.57	.08
*49	0	0	0	0	1	1	.25	.75	.33	.33	.57
50	1	0	0	0	1	1	.84	.16	.89	.33	.57
51	0	1	0	0	1	1	.05	.95	.08	.33	.57
*52	1	1	0	0	1	1	.47	.53	.57	.33	.57
53	0	0	1	0	1	1	.84	.16	.33	.89	.57
54	1	0	1	0	1	1	.99	.01	.89	.89	.57
55	0	1	1	0	1	1	.47	.53	.08	.89	.57
56	1	1	1	0	1	1	.93	.07	.57	.89	.57
57	0	0	0	1	1	1	.05	.95	.33	.08	.57
*58	1	0	0	1	1	1	.47	.53	.89	.08	.57
59	0	1	0	1	1	1	.01	.99	.08	.08	.57
60	1	1	0	1	1	1	.13	.87	.57	.08	.57
*61	0	0	1	1	1	1	.47	.53	.33	.57	.57
*62	1	0	1	1	1	1	.93	.07	.89	.57	.57
63	0	1	1	1	1	1	.13	.87	.08	.57	.57
*64	1	1	1	1	1	1	.70	.30	.57	.57	.57

5D Estimating achievement by scores S_i and S_g

In this appendix, it will be proved that the individual and group scores are unbiased estimators of achievement with a smaller variance than the sample proportion.

Assume that both $P(r_{i:k}|C_m)$ and $P(e_k|C_m)$ are fixed over trials. Results will be derived for a single cue profile C_m (so that the conditioning can be dropped), but results can be directly generalised to the general case with more cue profiles.

To investigate the properties of the sample proportion, define a random variable X_t, taking value 1 if $r_{it} = e_t$, and value 0 if $r_{it} \neq e_t$. This variable follows a Bernoulli distribution with $p = P(r_{i:1})P(e_1) + P(r_{i:-1})P(e_{-1})$ for the present case of dichotomous events and responses. Then $Z = \sum_{t=1}^{T} X_t$ follows a binomial distribution, with $E[Z] = Tp$ and $\text{var}(Z) = Tp(1-p)$. For the sample proportion $Z/T = \widehat{p}(r_i = e)$ of this binomial distribution, we have

$$E[Z/T] = P(e_1)P(r_{i:1}) + [1 - P(e_1)][1 - P(r_{i:1})] = A_i,$$

and

$$\begin{aligned}\text{var}(Z/T) &= \frac{p(1-p)}{T} \\ &= \frac{[2P(e_1)P(r_{i:1}) - P(e_1) - P(r_{i:1})]^2 + 2P(e_1)P(r_{i:1}) - P(e_1) - P(r_{i:1})}{T}.\end{aligned} \tag{5D.1}$$

To investigate the properties of the score variable S_i as defined in Equation 5.8, define a random variable r_{1t} taking value 1 for $r_{it:1}$ and 0 otherwise, and let $R_1 = \sum_t^T r_{1t}$ and $R_{-1} = T - R_1$. We have

$$\begin{aligned}E[S_i] &= P(e_1)E[R_1/T] + P(e_{-1})E[R_{-1}/T] \\ &= P(e_1)P(r_{i:1}) + [1 - P(e_1)][1 - P(r_{i:1})] = A_i,\end{aligned}$$

and

$$\text{var}(S_i) = \frac{P(r_{i:1})(1 - P(r_{i:1}))(2P(e_1) - 1)^2}{T}. \tag{5D.2}$$

By subtracting (5D.2) from (5D.1), it follows that

$$\text{var}(Z/T) = \text{var}(S_i) + P(e_1)[1 - P(e_1)].$$

Hence, $E[Z/T] = E[S_i] = A_i$, but $\text{var}(Z/T) \geq \text{var}(S_i)$.

5E Complete ecological systems in experiment 2

Table 5E contains the details of the two complete ecological systems that were used in experiment 2. The cues c_j are conditionally independent given the events e_k. For

each cue profile $C_m = (c_{1:x}, \ldots, c_{4:z})$, the corresponding probabilities $P(e_1|C_m)$ and $P(e_2|C_m)$ are given. Also, the probabilities $P(e_1|C_m^i)$ for the partial cue profiles $C_m^1 = (c_{1:x}, c_{2:y})$, $C_m^2 = (c_{1:x}, c_{3:y})$ and $C_m^1 = (c_{1:x}, c_{4:y})$ are given. Note that the values of $P(e_2|C_m^i)$ are not given, but they can easily be reconstructed from the information provided, since $P(e_2|C_m^i) = 1 - P(e_1|C_m^i)$.

Table 5E: Ecological system of experiment 2

m	c_1	c_2	c_3	c_4	$P(e_1\|C_m)$	$P(e_2\|C_m)$	$P(e_1\|C_m^1)$	$P(e_1\|C_m^2)$	$P(e_1\|C_m^3)$
					High-validity				
1	0	0	0	0	.67	.33	.33	.33	.33
2	1	0	0	0	.97	.03	.89	.89	.89
3	0	1	0	0	.25	.75	.08	.33	.33
4	1	1	0	0	.84	.16	.57	.89	.89
5	0	0	1	0	.25	.75	.33	.08	.33
6	1	0	1	0	.84	.16	.89	.57	.89
7	0	1	1	0	.05	.95	.08	.08	.33
8	1	1	1	0	.47	.53	.57	.57	.89
9	0	0	0	1	.25	.75	.33	.33	.08
10	1	0	0	1	.84	.16	.89	.89	.57
11	0	1	0	1	.05	.95	.08	.33	.08
12	1	1	0	1	.47	.53	.57	.89	.57
13	0	0	1	1	.05	.95	.33	.08	.08
14	1	0	1	1	.47	.53	.89	.57	.57
15	0	1	1	1	.01	.99	.08	.08	.08
16	1	1	1	1	.13	.87	.57	.57	.57
					Low-validity				
1	0	0	0	0	.03	.97	.33	.33	.33
2	1	0	0	0	.01	.99	.08	.08	.08
3	0	1	0	0	.33	.67	.89	.33	.33
4	1	1	0	0	.08	.92	.57	.08	.08
5	0	0	1	0	.33	.67	.33	.89	.33
6	1	0	1	0	.08	.92	.08	.57	.08
7	0	1	1	0	.89	.11	.89	.89	.33
8	1	1	1	0	.57	.43	.57	.57	.08
9	0	0	0	1	.33	.67	.33	.33	.89
10	1	0	0	1	.08	.92	.08	.08	.57
11	0	1	0	1	.89	.11	.89	.33	.89
12	1	1	0	1	.57	.43	.57	.08	.57
13	0	0	1	1	.89	.11	.33	.89	.89
14	1	0	1	1	.57	.43	.08	.57	.57
15	0	1	1	1	.99	.01	.89	.89	.89
16	1	1	1	1	.96	.04	.57	.57	.57

5F Cue utilisation in group tasks in experiment 2

Table 5F contains all cue utilisation coefficients υ_j for each group in experiment 2. Utilisation coefficients υ_j with a single subscript refer to the utilisation of separate cues c_j. Utilisation coefficients with more subscripts refer to the utilisation of (partial) cue profiles. Besides the utilisation coefficients, the consistency coefficient ξ is given for each group.

Table 5F: Cue utilisation coefficients for group responses in experiment 2

	ξ	v_1	v_2	v_3	v_4	v_{12}	v_{13}	v_{14}	v_{23}	v_{24}	v_{34}	v_{123}	v_{124}	v_{134}	v_{234}	v_{1234}
								High-validity								
1	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2	.41	.00	.07	.03	.00	.03	.00	.00	.01	.08	.00	.02	.02	.00	.02	.11
3	.93	.83	.00	.00	.00	.03	.03	.03	.00	.00	.00	.00	.00	.00	.00	.00
4	.81	.56	.00	.03	.03	.03	.08	.08	.00	.00	.00	.00	.00	.01	.00	.00
5	.88	.46	.05	.05	.01	.09	.09	.06	.00	.01	.01	.01	.00	.00	.04	.00
6	.61	.19	.01	.00	.19	.00	.02	.13	.00	.00	.02	.03	.00	−.01	.03	−.01
7	.74	.31	.11	.05	.01	.12	.05	.02	.01	.02	.00	.02	.00	.01	.01	.00
8	.88	.30	.10	.01	.10	.12	.01	.12	.00	.02	.00	.00	.04	.00	.05	−.01
9	.86	.72	.01	.01	.00	.06	.06	.00	.01	.00	.00	.00	.00	.00	.00	.00
10	.88	.66	.01	.01	.00	.05	.05	.07	.00	.01	.01	.00	−.01	−.01	.00	.00
11	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
12	.31	.07	.03	.00	.03	.01	.00	.04	.00	.00	.00	.00	.02	.05	.00	.05
13	.81	.56	.00	.03	.03	.03	.08	.08	.00	.00	.00	.00	.00	.01	.00	.00
14	.81	.56	.03	.00	.03	.08	.03	.08	.00	.00	.00	.00	.01	.00	.00	.00
15	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
								Low-validity								
16	.75	.01	.10	.10	.19	.00	.06	.02	.02	.08	.08	.05	.01	.01	.01	.00
17	.81	.56	.00	.00	.03	.03	.10	.08	.00	.00	.03	.00	.00	−.02	.00	.00
18	.74	.08	.16	.08	.03	.04	.02	.00	.04	.02	.00	.02	.03	.15	.03	.01
19	.93	.00	.00	.83	.00	.00	.03	.00	.03	.00	.03	.00	.00	.00	.00	.00
20	.93	.24	.03	.14	.07	.08	.12	.09	.04	.01	.09	−.01	.00	.01	.02	.00
21	.74	.11	.05	.19	.05	.06	.07	.01	.07	.01	.03	.00	.06	.01	.02	.02
22	.56	.00	.24	.03	.00	.01	.07	.00	.08	.01	.00	.08	.02	.00	.00	.00
23	.81	.00	.03	.38	.00	.00	.01	.07	.08	.03	.01	.07	.06	.10	.04	−.07
24	1.00	.19	.19	.00	.19	.12	.00	.12	.00	.12	.00	.00	.07	.00	.00	.00
25	.88	.46	.05	.05	.01	.09	.09	.06	.00	.01	.01	.01	.00	.00	.04	.00
26	.68	.03	.03	.24	.07	.00	.01	.04	.01	.01	.09	.04	.02	.00	.11	−.03
27	.43	.14	.00	.03	.07	.00	.01	.02	.03	.03	.01	.02	.05	.04	.00	−.02
28	.86	.00	.01	.72	.00	.00	.00	.01	.06	.00	.00	.00	.01	.06	.00	.00
29	.86	.47	.00	.05	.05	.04	.10	.10	.01	.01	.02	−.01	−.01	.01	.02	.00

6
The unfounded demand for consensus

A camel is a horse designed by committee.

Sir Alec Issigonis

The achievements of a scientific discipline are often read from the consensus between scientists in that particular discipline. For psychology, where consensus is hard to find, this practice leads to the somewhat depressing conclusion that not much has been achieved. Of course, there are theories, such as phrenology, that have been rejected by all psychologists. But a theory that has been generally accepted appears non-existent. Introductions to psychology offer a motley bunch of (mini-)theories that, if not mutually contradicting, show little interconnection. This leads to great despair in some first-year students and to qualifications of incoherence and pseudoscience by non-psychologists. For others, it is a sign of psychology's scope and vitality. Whatever one's reaction, psychology contrasts sharply with a field such as physics, which does show consensus and coherence.

According to a classical view, scientific consensus is guaranteed by use of the Scientific Method, which because of its access to the Truth, demands unanimous agreement between the scientists that follow it (Laudan, 1984). Unfortunately, such a naive view of science is untenable. It has now been commonly accepted – at least in the philosophy of science – that there are no compelling criteria to distinguish knowledge, truth and science from their counterparts. In this respect, physics is no different than psychology; both the rejection of phrenology and the acceptance of Einstein's theory of special relativity do not stem from conclusive evidence for or against the theory. This much should be clear from the previous discussion of underdetermination and the Duhem-Quine thesis (see Chapter 2). But if all theories are underdetermined and theory-choice unjustified, a certain arbitrariness seems unavoidable. Is it then at all possible to defend science as a rational enterprise?

6.1 The consensus imperative

Ziman (1968) and De Groot (1971) solve this problem as follows: consensus is not the product of the infallible Scientific Method, it *is* the scientific method itself. The striving for consensus is the distinguishing characteristic of science and as such a norm for scientific action. This step, elaborated in the Forum Theory (De Groot, 1961, 1971, 1977, 1982), is tempting. Contrary to absolute certainty based on fact and logic, consensus is a viable goal. But upon a closer inspection, there is little justification for the consensus imperative. That there are scientists who strive for consensus is an empirical statement, which is either true or false. But that all scientists should strive for consensus is a normative statement, and truth has no say in such a matter. Unless the categorical imperative is turned into a hypothetical one. Such a goal-directed version would be 'If you want to achieve X, then you should reach consensus on matter Y'. This proposition is true if consensus on Y is necessary for the attainment of X, or if it makes the attainment of X easier. One defining characteristic of consensus is unanimity in opinion and evaluation. In science, opinions are usually generalising statements about the world, and evaluations the truth-value of such statements. Hence, we may replace Y with an opinion or its evaluation. But what about X? What scientific goal demands consensus?

From the consensus-imperative, De Groot wanted to arrive at 'a *normative minimal methodology of the empirical sciences*, such that, in addition, an "alternative methodology" is impossible'[1](De Groot, 1971, p.8). One way to look at normative methodology is as a system of rules and heuristics for the scientific evaluation of knowledge claims. In order for consensus to function as the foundation of such a methodology, it should function as a scientific goal, means or criterion. According to De Groot (1971), consensus meets this demand entirely: unanimous judgement is goal, criterion and procedural prescription. But as will be argued here, consensus does not fulfill any of these roles: consensus is neither an adequate scientific goal, means, nor criterion.

6.2 Consensus as goal

The first question is whether consensus can be taken as a scientific means, or just a goal. I say 'just', because the attainment of consensus is far removed from classical goals such as the acquisition of knowledge and the formulation of true theories. Unless knowledge and truth are defined in terms of consensus. This last option appears a way out for adherents to the consensus-imperative, for as will become clear later, the view that consensus is an adequate scientific means is difficult to maintain. But what is the use of a consensus-imperative if consensus itself is the goal? A proposition such as 'If you want to achieve consensus, you should (attempt to) achieve consensus' is consistent, but not very informative. Of course, other imperatives are possible. As De Groot puts it:

> If one wants to be able to reach agreement – and after all, Forum-discussions

[1]Translated from Dutch by M.S. Original: 'een *normatieve minimale methodologie van de empirische wetenschap*, zodanig, dat er daarnaast géén "alternatieve methodologie" mogelijk is.'

> are about reaching agreement, about reaching a 'rational intellectual consensus' – then one has to "report *objectively*, reason *logically* and explain oneself *clearly*, and then *honesty* prevails"; and it is clear that one *cannot* reach agreement on "*vague*, *inconsistent*, or otherwise *unfalsifiable* statements"(1971, p.7)[2].

Taking it as a given that science strives for consensus, De Groot derives the 'rules of the game' that will make consensus possible. But the derivation is overstated, of course. For instance, there is consensus amongst psycho-analysts about the causal relation between unconscious drives and behaviour. Since this presumption usually doesn't lead to testable predictions, a demand such as falsifiability is not a prerequisite for consensus. Moreover, if consensus is the only goal, any means that results in consensus should suffice. In this respect, convincingly lying and carefully adjusting the data might be better means than honest presentation. But such strategies do not harmonise with the usual scientific mores.

6.2.1 Rational consensus

Apparently, not every consensus suffices. Ziman and De Groot's goal is a *rational* intellectual consensus. This specific species of agreement is difficult to define. De Groot (1982, p.245) states that the behaviour of a problem-solver is rational only if it suits the goal, in light of the specific structure of the problem, or the 'problem environment'. But this common usage of 'rational' is not applicable to consensus as a goal. For a goal always suits itself, so that every consensus should be considered rational. Putting a priori restrictions on the subject of consensus, the proposition or evaluation, is not an option either. The problems with such restrictions were the main impetus for the consensus-imperative in the first place! Letting the definition of rational consensus itself be a matter of consensus merely transfers the problem, for one cannot say whether *that* consensus is rational or not.

We could speak of rational consensus when it is the product of rational behaviour. All individuals in a group behave functionally with respect to their goals and the end-result is consensus. This condition implies that, besides consensus, at least one other goal is aspired. Consensus is rational if it fits with each one's goals in the light of the problem or problem environment. Clearly, such a consensus cannot always be attained whenever the individual goals are different or conflicting. But a compromise may allow each one's aims to be partly met. Whether such a compromise is rational depends on the ranking of the goals. Is consensus necessary, or is consensus more important than the gratification of other goals? Matters of policy often require a decision at a specific time. Moreover, in order to secure implementation, this decision has to be supported by as many parties as possible. For such matters, reaching consensus is crucial. But on what grounds can one pinpoint a time by which, for the whole of science, it has to be decided for instance that the development of children

[2]Translated from Dutch by M.S. Original: 'Wil men het eens kùnnen worden - en in forumdiscussies gaat het er immers om het eens te worden, gaat het om 'a rational intellectual consensus' – dan mòet men wel "*objectief* rapporteren, *logisch* redeneren en *helder* uitleggen, en dan duurt *eerlijk* het langst"; en het is duidelijk dat men het over "*vage*, *inconsistente* of anderszins *niet-falsificieerbare* beweringen" *niet* eens kàn worden.'

is a stage-wise, rather than continuous process? Of course, science is not devoid of issues of policy, such as the allocation of grants to research-projects. Such decisions certainly affect the content of science. Without grants, it may not be possible to collect empirical evidence. But they do not affect the truth-value of a theory, or at least should not. Goals are a necessary aspect of the definition of rationality, but are not themselves subject to evaluations of rationality. Holding a consensus goal is neither rational nor irrational. But when different goals are held concurrently, the act of striving for consensus does not escape evaluation in terms of rationality. If the attempt to reach consensus impedes the meeting of other aims, and these other aims are deemed more important than consensus, the gratification of the consensus goal is irrational. I find it hard to imagine that, when it comes to scientific statements and their evaluations, consensus is necessary or more important than other cognitive aims such as empirical adequacy, consistency, or truth. Actually, that consensus is subordinate to at least one other goal follows from the need to distinguish arbitrary from rational consensus. The rationality of concerned parties is thus of primary importance. If the condition of rationality necessarily results in consensus, then the consensus goal is parasitic on the rationality goal and has no additional value. If this is not the case, then the question rises whether it is possible to reach agreement without abandoning rationality. If rationality is taken as a form of optimising, as is usual, this will often be impossible. For if actions are not axiologically underdetermined, then there is only one rational act for each individual. In order to reach agreement, some parties will have to give up this rational act. Consensus and rationality thus appear as conflicting, but this is not necessarily the case. The idea is of course that information is shared in the process of reaching agreement. So the informational base on which to found judgement is extended. In light of the new information, earlier judgements may no longer be rational. If the shared information allows for only one rational judgement, than a rational consensus is reached. But now the consensus goal is again parasitic, for it is not reaching agreement that matters, but forming a judgement on the basis of all available and relevant information. Moreover, reaching consensus is now more a means for gathering information than a goal itself. More about consensus as a means later. As a scientific goal, consensus plays at most a subordinate role.

6.2.2 Social constructs

In the previous section it was argued that arbitrary consensus is certainly not a scientific goal, while rational consensus is either impossible, or depends so strongly on the rationality of the concerned parties, that is it redundant. We could now stop discussing the role of consensus as scientific goal. However, consensus may become a goal in a different way, namely by practical equivalence to another aim such as knowledge or truth. This is the position of social constructivists such as Gergen (1985), who argue that the products of science are the result of social processes such as persuasion and negotiation. Social constructivism is not a well-defined position, and many distinctions are possible (Kukla, 2000). Typical is the thesis that scientific knowledge consists of interpretations formed through social interaction. In a way, this is true. Scientific theories are not discovered, but created, and scientists do not work in a social vacuum. In this respect, scientific theories are social constructions. But

this thesis is as trivial as it is true. The social constructivist thesis is interesting only if its claims extend further.

As pointed out by Hacking (1999), the thesis that X is a social construct is based on the thesis that

> X need not have existed, or need not be at all as it is. X, or X as it is at present, is not determined by the nature of things; it is not inevitable. (p.6)

According to social constructivists, the phenomena studied by scientists are not inevitable manifestations of objectively existing entities and processes, but post hoc, in social interaction construed theoretical entities (Nelson, 1994). Radical social constructivists couple this with a relativistic viewpoint in which every possible consensus is on a par. The agreement to call something else a 'fact' would have maybe led to a totally different, but just as coherent world-view as another. Every aspect of science is negotiable, and the result of negotiation is not determined by objective reality. Not only are theories social constructs, but so are the phenomena they intend to describe. Underlying social constructivism is a consensus theory of truth: the truth of a scientific proposition or theory consists only of their being labelled 'true' by members of the scientific community (Fine, 1996). The consensus theory of truth will be discussed later. What is important here is that the theory takes truth as relative to the group in which consensus exists. Most empirical scientists react with reservations to such relativism. Reaching agreement, without further qualifications, is of course no scientific goal. Social constructivism's cold reception confirms this.

Not all constructivists are radical relativists. According to moderate constructivists, such as Knorr-Cetina (1995) and Liebrucks (2001), the material world gives a certain resistance which places restrictions on the social constructions that attempt to describe her. What they contest is the idea that scientific theories are linguistically objective representations of reality (Knorr-Cetina, 1995). Who they contest, however, is unclear. Are there any scientists who entertain this naive idea? The strength of the social constructivists does not lie in their reuse of philosophical arguments, but in their observational studies of scientific practice. If we are allowed to conclude anything from these (as some social constructivists would contest), then these show how many subjective choices must be made. Every test of a theory rests on subjective decisions concerning the specific form of the manipulation, the measurement-model, the required amount of observations, etc. Because these choices are not determined by the nature of the phenomenon under study, the acceptance of a test, and with that, of the theory which is subjected to it, is more or less a matter of convention. This is not a new idea. As mentioned in Chapter 2, Poincaré (1905/1979), the logical positivists, Popper (1959), and Kuhn (1970), all defended a particular form of conventionalism. But while they explain the (implicit) choice for a particular convention by internal factors such as the usefulness, simplicity, or coherence of the convention, the social constructivists call upon external factors such as shared interests and stakes of particular scientific communities. Of course, there are cases where conventions were directed by social factors. Danziger (1990) quite convincingly argues how the move from psychology as the study of the individual to psychology as the study of aggregates was influenced by the success of the latter type of research in educational and

		B	
		b_1	b_2
A	a_1	α, β	$-\alpha, -\beta$
	a_2	$-\alpha, -\beta$	α, β

Figure 6.1: A simple coordination game

military circles. But, as Danziger himself also admits, this finding in no sense indicates the necessity or desirability of the social determination of scientific knowledge. Since the conclusion that scientific facts are social constructions is often not taken to be reflexive – that is, applicable to itself – social constructivists neither take scientific facts necessarily as social constructions (Kukla, 2000). Neither the necessity nor the desirability of the social constructivists picture of science has been convincingly argued for. As for necessity, the relativistic social constructivists have put themselves in an awkward position. According to their own view, their view is but one amongst equals. Another paradox of self-reference. While scientific theories are, in a trivial sense, social constructions, the genesis of scientific theories has no inevitable effect on the relation between a theory and the reality which it aims to describe. The distinction between the context of discovery and the context of justification can be usefully applied here. As for desirability, or rather undesirability, I find the cold reception of social constructivism amongst empirical scientists (and non-social-constructivist philosophers of science) typical. Social constructivism is an empirical theory that claims how science actually proceeds, rather than a normative theory that claims how it ideally *should* proceed.

6.2.3 Coordination games

If consensus is purely a goal, then prohibiting consensus is just as nonsensical as demanding it. There are no rational grounds on which to recommend or dissuade a goal. But I dare to contest that scientists predominantly strive for consensus. Consensus is a goal in problems of interpersonal coordination. Such coordination problems arise when conditional preferences are held of the kind 'I prefer to do X if, and only if, you do X also'. To stay within the decision-theoretic framework of Chapter 2, the situation will be described as a game. A coordination game is characterised by symmetrical conditional preferences. In such a case, coordination of the action X is necessary for a preferable outcome. A simple example of a coordination game is given in Figure 6.1. Both A and B have two available actions, a_1 and a_2 for A, and b_1 and b_2 for B. For instance, a_1 and b_1 might be 'accept theory T_1' and a_2 and b_2 'accept theory T_2'. In this game, consensus in cognitive action is the only acceptable outcome for both players. Given that both A and B coordinate their actions, the choice for either (a_1, b_1) or (a_2, b_2) is entirely arbitrary. It does not matter on which action agreement is reached, as long as agreement is reached. Such a situation might be palatable for radical social constructivists, but most scientists will find it rather distasteful. Clearly, an important player has been left out: nature. Suppose that T_1 describes nature to be in state θ_1, and T_2 describes nature to be in state θ_2. Then

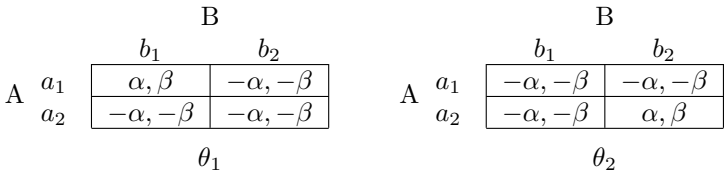


Figure 6.2: A three player coordination game

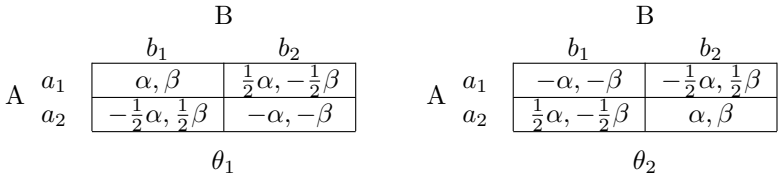


Figure 6.3: Amended three player coordination game

the three-player game might be something like the one in Figure 6.2. The conditional preferences in this game are more plausible. Conditional upon θ_1, both players prefer (a_1, b_1) to any other outcome, while conditional upon θ_2, both players prefer (a_2, b_2) to any other outcome. In this game, the only acceptable outcomes are those in which all three players coordinate their action. Of course, the coordination should now not be taken literally, for the idea that nature coordinates her state to match the beliefs of scientists is a silly one. Rather, both A and B need to coordinate their actions with nature and each other. But the situation in Figure 6.2 is still somewhat unrealistic. Surely, A would find outcome (a_1, b_2, θ_1) preferable to (a_2, b_2, θ_1). Even B, if not entirely egocentric, might prefer the former outcome to the latter. A more likely situation is one in which the epistemic utility of A conforms to the following ordering: $u_A(a_1, b_1, \theta_1) \succ u_A(a_1, b_2, \theta_1) \succ u_A(a_2, b_1, \theta_1) \succeq u_A(a_2, b_2, \theta_1)$. For such a preference structure, and similar ones for θ_2 and B, the game might look something like that in Figure 6.3. This game looks quite different from that in Figure 6.1. It is no longer a pure coordination game. One should realise that the epistemic utilities of the actions are given for certain situations, i.e. 'If the true state of nature is θ_1, then the utility of a_1, conditional upon B taking action b_1, is α'. Assume for the moment that there is no uncertainty as to the true state of nature, A knows for certain that it is θ_1. Why would A assign a higher utility to (a_1, b_1, θ_1) than to (a_1, b_2, θ_1)? In both outcomes, he adopts a true belief. Why should B adopting a false belief diminish A's utility of adopting a true belief? In an altruistic mode, A might be concerned with the well-being of B, which is negatively affected by adopting false beliefs. In a more egocentric mode, A might be concerned with his own well-being. If B has decisive power over A publishing an article exposing his belief, and B will only allow publication when she agrees with A, then surely $u_A(a_1, b_1, \theta_1) \succ u_A(a_1, b_2, \theta_1)$. However, if A is solely concerned with publishing, then his utilities would be better reflected by the pure coordination game depicted in Figure 6.1. To assure publication, A would simply

need to assure that his action conforms to B's action. But this offers a perverse view of scientific behaviour. Moreover, agreement is then not a goal itself, but a means to the goal of publishing. But insofar as A has a goal of adopting true beliefs and an additional goal of publishing, the situation in Figure 6.3 is plausible (although consensus would still not be a goal in itself). And what about social validation, shouldn't agreement between A and B raise the validity of an action? Maybe, but then again, consensus is not a goal, but a criterion of validity.

If an individual has a goal of consensus, then everything else being equal, he should assign a higher utility to an outcome in which there is consensus than an outcome in which there is dissensus. This is so for A and B in the game in Figure 6.1, but not in the games in Figure 6.2 and Figure 6.3. For although $u_A(b_i|a_i, \theta_i) \succ u_A(b_j|a_i, \theta_i)$ in both Figure 6.2 and Figure 6.3, so that A prefers agreement with B if his belief agrees with the state of nature, $u_A(b_j|a_j, \theta_i) = u_A(b_i|a_j, \theta_i)$ for $j \neq i$ in Figure 6.2, and $u_A(b_j|a_j, \theta_i) \prec u_A(b_i|a_j, \theta_i)$ for $j \neq i$ in Figure 6.3. In words, these two latter games describe a situation in which, conditional upon achieving the goal of a true belief, consensus is an additional goal, but conditional upon not achieving the goal of a true belief, consensus is not an additional goal. This means that, insofar as the orderings of the utilities in these two games conform to those entertained by actual scientists, consensus is a subordinate goal, only aspired after the realisation of other goals.

Although De Groot and Ziman assigned too much status to consensus, they were right in one respect: scientific knowledge is public knowledge. This is to say that scientists should not only strive for true or otherwise adequate beliefs, but they should also strive to make these knowable and intelligible for others, who may employ or criticise them as they see fit. Science is a social activity in which mutual exchange and criticism are of fundamental importance. But such a (justified) demand for publicity is something completely different to a demand for consensus. Consensus is not a primary scientific goal.

6.3 Consensus as means

If consensus is not a goal itself, then the consensus-imperative has foundation only when consensus is a means to some other goal. De Groot (1961, 1971) and Ziman (1968) do not show relativistic sympathies and would not consider themselves social constructivists. They do not deny that truth is the quintessential scientific goal, but the problem is there are no adequate means available to reach this goal. The methodological rules that are employed are not compelling rules by which truth is discovered. In order to justify these rules as goal-directed means, De Groot and Ziman replace the goal of truth by rational consensus. This consensus is not arbitrary agreement, but one that approaches truth as closely as possible. Thus, reaching consensus is considered a means to discover truth.

If consensus is a means, then it should result in something. As mentioned in Chapter 2, consensus is usually taken as a prerequisite for interpersonal coordination. Let's go back to the simple coordination game in Figure 6.1. The problem usually addressed in game theory is how such coordination problems can be solved without

direct communication between A and B. In order to reach either preferable outcome without communication, consensus is required. The situation is similar to the problem of coordinated attack described in Chapter 2. A will decide on a_1 only if he believes that B will decide on b_1, while B will decide on b_1 only if she believes A will decide on a_1. So A will expect B to decide b_1 if he believes that B believes that he will decide a_1 (i.e. $B_A B_B B_A a_1$), etc. Again, coordination without communication requires common belief $E^{\infty}_{\mathcal{P}}$. In this case, the outcome of consensus requires that this consensus pre-existed in the minds of both players. But a situation without possible communication is not one encountered in scientific practice. Moreover, in such coordination problems, consensus is a means to achieve consensus. But if consensus is not an important scientific goal, why attempt to achieve it?

The consensus-imperative is justified insofar as consensus is a means to another goal. Consensus should at least have an additional result to consensus. Different effects of consensus have been mentioned. For instance, members of a group with a high level of consensus have a greater sense of well-being (Bliese & Halverson, 1998). In social psychology, much research has been conducted on the differences between small groups deciding by majority and those deciding by unanimity. Miller (1989) gives an overview of the research on this subject and concludes that a unanimity-rule results in

- a higher probability that no final decision is reached, but if reached, it is more often a compromise in which minority opinion is better reflected,
- a higher correspondence amongst individual preferences, i.e. more consensus,
- a longer discussion preceding the decision which is perceived as more uneasy, difficult and conflict full, but also as more thorough and adequate,
- group members judging each other as more likeable.

A unanimity-rule is also judged as fairer than a majority-rule, probably because it results in more consensus, so that the group decision corresponds more to individual decisions. Finally, consensus results in a higher trust in a decision or opinion (Orive, 1988). This can be explained by Festinger's (1950, 1954) theory of social comparison (see Chapter 4). According to this theory, the judged validity of an opinion is positively related to the number of individuals holding that opinion. The more individuals agree on an opinion, the more valid that opinion will be judged.

6.3.1 The quality of group decisions

It would be nice if groups deciding under a unanimity-rule are justified in putting greater trust in their decisions. However, the results of research on the effect of a unanimity-rule on the quality of decisions are mixed. Holloman and Hendrick (1972) and Bower (1965) found group decisions under unanimity to be better than those arrived at by a majority-rule. Sorkin, West, and Robinson (1998) found that groups perform worse as the decision-rule becomes more strict, with a simple majority rule as least strict, and a unanimity-rule as most strict. Stasson, Kameda, Parks, and Zimmerman (1991) found a better performance in mathematical problem-solving for

groups under a simple majority rule. But, in those groups in which none of the members individually knew the solution, groups under unanimity arrived at the correct solution more often than groups under majority. The authors explain this last result by the more thorough discussion that took place in the former groups. This was also the explanation given by Holloman and Hendrick (1972) and Bower (1965). In this respect, the research of Postmes, Spears, and Cihangir (2001) is interesting. They compared groups working under a consensus norm to groups working under a norm of critical discussion. The quality of decisions was much higher in the last than in the former groups. Comparable results are usually found when comparing dialectical inquiry, devils advocacy, consensus and expert discussion methods. There is some disagreement whether dialectical inquiry or devils advocacy has better results, but both perform better than the other methods (Katzenstein, 1996). When a norm of critical discussion is not explicitly enforced, the tendency towards social harmony can disrupt the critical examination of possible solutions, as in the groupthink phenomenon (Janis, 1972). This happens when groups seek consensus prematurely, resulting in the gathering of too little information, alternative solutions and sources of possible failure. The potential of groups, stemming from the large informational base they can access, is thus unrealised. This is a general finding. Rather than focussing on the unique information each member can contribute, groups focus on the information which is shared by a large number of group members (Wittenbaum & Park, 2001). Furthermore, research on brainstorming shows that interactive groups produce less alternatives than so-called nominal groups, consisting of individuals working alone (Mullen, Johnson, & Salas, 1991). Different explanations have been given for this finding. For instance, there is motivation-loss resulting from working in a group, because there is no individual responsibility ('social loafing') and it is tempting to let others do the work (the 'free rider' effect). Also, some group members experience a certain inhibition stemming from fear of a negative evaluation by other members. Finally, procedural aspects also play a role, since group members must take turns in expressing their alternatives. The intuitive idea that groups arrive at better decisions because of mutual information exchange is not clearly supported by empirical research. The many experiments in the Asch-paradigm (see Chapter 4 and Levine & Thompson, 1996) show that consensus often leads to conformity to the consensual position.

To improve the quality of group decisions, group members must be persuaded to share as much relevant information as possible. Moreover, measures must be taken to eliminate such processes as conformity, social pressure, social loafing and free riding. A norm of critical dissensus appears a better means for this than one of consensus.

6.3.2 Individual vs group

The majority of the above mentioned research compares the performance of different kinds of groups, where it was expected from these groups that they arrived at a single decision or judgement. But how does the performance of groups compare to that of individuals? This question does not have an unambiguous answer. Which individual should be compared to the group, and how?

Usually, group judgements are compared to the judgements of the group members

before group discussion. In this case, there are at least three possible comparisons. The first one is between the group judgement and the judgement of the least competent member. Here, group judgement will most likely compare favourably to individual judgement. The comparison between the group judgement and the average individual judgement will be less favourable to the group judgement, but the group judgement is often better than the average individual judgement (Hill, 1982; Laughlin, Bonner, & Miner, 2002). When comparing the group judgement to the judgement of the most competent individual, groups usually perform worse (Hill, 1982; Stasson et al., 1991). In the two experiments reported in Chapter 5, the accuracy of collective decisions lay between the accuracy of the average and best member's decisions. Statistical tests showed that the accuracy of the collective decisions was not significantly different from the accuracy of the decisions of the most competent group member. In addition to this finding, the results of these experiments indicated that the relative accuracy of a group compared to its best member depends on how information is distributed amongst the group members. In the case that information is completely distributed, the group as a whole can base its decision on more information than any individual alone. As such, the potential accuracy of a group is higher than the potential accuracy of any individual alone, no matter how competent. Groups were not able to realise their potential. Hence, even if one has good reasons to expect the group to perform better than its best member, this expectation is not generally confirmed.

Suppose that someone must choose between a group or an individual judgement. If the group judgement is always better than the individual judgement of the least competent individual, choosing the group judgement is a security-strategy. The worst possible outcome of the group judgement is at least as good as the worst possible outcome of an individual judgement. Since group judgements are predominantly better than the average individual judgement, the expected outcome of the group judgement is also preferable to the expected outcome when randomly choosing the judgement of an individual. But, since the group judgement is usually no better, and often worse than that of the most competent individual, choosing the group judgement is certainly not a strictly dominant strategy. If there is information regarding the competence of individuals, the judgement of the nost competent individual is preferable to the judgement of the group.

When comparing individual judgements to group judgements, one should not only ask 'which is better?', but also, 'better for who?'. When the group judgement is not at least as good as the individual judgement of any of its members, then the group process has a detrimental effect on at least one member. The question is then whether this is justifiable by the possible positive effect on other group members. When reaching consensus is viewed as a learning process, the group judgement should be internalised by the group members. In this case, the majority of group members will learn, but the remaining part will unlearn. Weighing collective gains and individual loss is principally an ethical question. I can not see how a Scientific Law that imposes a particular weighing onto all scientists can be defended.

6.4 Consensus as criterion

In the previous, I argued that consensus is not an adequate scientific goal or means. What remains for consensus is a role as epistemic criterion. We should distinguish between consensus as a definition of truth, and consensus as criterion of an otherwise defined truth, for instance in terms of correspondence to states-of-affairs. The consensus definition of truth has serious drawbacks. Van Heerden (1980) mentions three. Firstly, it is parasitic on another definition of truth, since the presence of unanimous agreement is an empirical matter. One could reply that the presence of consensus can be made a matter of unanimous agreement itself, but this would lead to an infinite regress. Secondly, a statement cannot in retrospect be characterised as true in the time preceding unanimous agreement. This conflicts with the time-less character of truth. It should be noted that Van Heerden's objections hold only for a definition of truth in terms of actual consensus, not for one in terms of counterfactual consensus. According to Habermas' consensus theory of truth (see Hesse, 1980), a statement is true if competent actors in an ideal speech-situation would unanimously agree on it. The conditions for an ideal speech situation are

(a) every person with the competence to speak and act can participate.

(b) everyone is allowed to contribute and criticise any assertion, and can express any attitude, wish or need.

(c) no one may be prevented, either by internal or external pressure, to exert his or her rights as given in (a) and (b).

Of course, the ideal speech situation is just this: an ideal. In practice, it will not be realised and Habermas' theory does not provide a practical criterion of truth. Moreover, if the definition of truth is consistent, then for any set of contradicting statements, there must be consensus on one (and only one) element. In the ideal speech situation, there must always be a winning argument for one. An example:

I think it is safe to assume there is consensus that the statement 'either aliens populated Mars, or they did not' is true. Aliens did not populate Mars 'a little', or 'approximately'. If we denote the statement 'Aliens populated Mars' as X, and the statement 'Aliens did not populate Mars' as $\neg X$, then '$X \vee \neg X$' is true. But if unanimous agreement cannot be reached on X or $\neg X$, then both are false. So $\neg X \wedge \neg(\neg X) =$ '$\neg X \wedge X$, which is a logical contradiction and moreover contradicting the true statement '$X \vee \neg X$'[3]

Finally, I should mention Van Heerden's (1980) third objection. With a consensus definition of truth, reaching consensus on probability, deficient English, or ambiguity, results in the indistinguishability of probability, deficient English, and ambiguity. For if truth can be defined solely in terms of agreement, then so can other aspects of statements. Why would truth be the only attribute definable in terms of consensus?

[3]The second contradiction may seem superfluous in light of the first, but it isn't. For an adherent to the consensus theory of truth might not object to the logical contradiction $\neg X \wedge X$, for the contradiction rests on the 'law of the excluded middle', which may not be unanimously accepted. However, he should be persuaded by the inconsistency between $\neg X \wedge X$ and $X \vee \neg X$, since both are true according to his own definition.

If consensus is the only criterion, these different attributes are indistinguishable after consensus is reached. In order to do this, other norms must be introduced, but then consensus is deprived of its decisive role. On the basis of these conclusions, even the ideal consensus theory of truth should be rejected.

Although consensus does not provide a good definition of truth, maybe it could function as a (fallible) criterion of truth otherwise defined. This seems to be the idea behind the statement 'true is (for the time being) that which the forum ... holds true.'[4](De Groot, 1982, p.248). It corresponds to the social constructivist thesis that truth consists only of the assignment of the label 'true' by a community. As an empirical observation, this is correct. The only observable correspondence between all statements that are held true, is just that they are held true. But truth is of course not an empirically observable property. Furthermore, I don't think there are many scientists who assume that everything they hold true is actually true. An important function of the concept of truth is to distinguish between true and false beliefs. An adequate definition of truth should at least allow for mistaken beliefs. Having a belief should not make it true by definition. Since this does not hold for the definition above, we should see consensus as an epistemic criterion, and not a definition of truth. But, for difficult problems, consensus is not epistemically informative, as the argument in Appendix 6A shows. Moreover, consensus is a rather peculiar epistemic criterion. Any influence it exerts is self-reinforcing. Anyone who is persuaded by consensus to adopt a certain viewpoint strengthens the consensus. As more people hold a view, there is more justification for the view, and more people will be persuaded to adopt the view, etc. In the end, there is not much left of the indicative function that consensus may have served initially. In order to function properly, a consensus criterion should not actually be applied. Thus, consensus is not a useful criterion.

6.5 Cooperation without consensus

Consensus is not a goal, means, or useful criterion in the empirical sciences. It is possible that a theory which is true, leads to the best predictions, or is the most useful, will also be unanimously adhered to. But this should not be confused with the idea that if a theory is unanimously adhered to, it is also true, leads to the best predictions, or is the most useful. This is of course the fallacy of confirming the consequence ('if true then consensus', 'consensus' thus 'true'). And even if consensus is always accompanied by truth, then it is a naturally occurring consensus. By enforcing a consensus imperative, consensus becomes a manipulation, with an unclear validity. Demanding consensus should result in more consensus than without the demand. Otherwise the imperative is inconsequential. If consensus arises naturally when persons cannot find any reasonable points of disagreement, then the surplus of manipulated consensus consists of fictitious agreement. Again, by striving for consensus, it loses its indicative function.

Subjecting a belief to evaluation by other scientists mainly serves the purpose of detecting errors. It is because there are no compelling rules for the evaluation

[4]Translated from Dutch by M.S. Original: 'waar is (voorlopig) dat wat het forum ... voor waar houdt'.

of knowledge claims, and because it is difficult for a scientist to rise above his or her theoretical framework, that the opinions of others are informative. Consensus regarding a judgement means that, for the time being, no grave errors have been detected. The validity of such a social tests critically depends on the possibility of detecting errors. If certain errors are due to certain presumptions that are part of a theoretical framework, this possibility is limited when others share the framework. The more agreement there is beforehand, the more people share the same theoretical framework, the less valid the test. Consensus as epistemic criterion is availed by dissensus.

It cannot be denied that science is, and should be, a social enterprise. Mutual sharing of empirical data, theories, and ideas, prevents that each scientists must reinvent the wheel. It is important that the ideas and observations are available and useable: scientific knowledge is public knowledge. Availability has its own problems, in which consensus plays a minor role, for instance when consensus between peer reviewers is a condition for publication in a scientific journal. Here, consensus has a limiting function, restricting the amount of public data and theory. One could argue that the usefulness of data and theories requires a certain level of consensus. According to Kuhn (1970), different paradigms are incommensurable, so that the meaning of an observation depends on the paradigm one adopts. Psychology is often taken as a pre-paradigmatic science. Some find this disturbing, maybe because Kuhn only speaks of 'normal science' in paradigmatic stages. Others regard Kuhn's normal science as rather pathological (Popper, 1974). A pre-paradigmatic science is characterised by the simultaneous existence of multiple (mini-)paradigms. If paradigms are truly incommensurable, then communication between scientists might seem impossible. But, that theoretical terms have a different meaning in different paradigms does not mean that it is impossible to understand both meanings (Laudan, 1996). Incommensurability does not mean that empirical data are unusable or meaningless.

According to De Groot (1977, 1990a), psychology has many schisms, apparent contrasts that arise for instance when there is no agreement on the right methods or meaning of basic terms. Reaching consensus is thus regarded as a medicine for the fragmentation of psychology. This is akin to the 'Kuhnian medicine' as Feyerabend (1974) describes it: 'restrict criticism, to reduce the number of comprehensive theories to one, and to create a normal science that has this one theory as its paradigm' (p.198). This was obviously not what Kuhn meant:

> I claim no therapy to assist the transformation of a proto-science to a science, nor do I suppose anything of this sort is to be had. If certain social scientists take from me the view that they can improve the status of their field by first legislating agreement on fundamentals and then turning to puzzle-solving, they are misconstruing my point (Kuhn, 1974, p. 245).

Consensus is not a medicine, but luckily scientific disagreement is not a disease which calls for one. While disagreement does indicate that there is 'something wrong' with the available theories – none is immediately acceptable – it does not indicate that there is something wrong with adherents to the different theories. As long as they keep asking themselves why they adhere to a particular theory and express the underlying reasons to themselves and others. A well-considered choice is only possible

to the extent that all conceivable advantages and disadvantages of a theory are known. The formulation of new theories will also be availed by such a thorough analysis. It is very likely that it is impossible to discover all advantages and disadvantages of a theory, but as a regulative ideal it certainly is not inappropriate. Aiming towards this ideal seems more availed by a norm of rational disagreement than one of agreement.

For good reasons, science is a social enterprise. From this idea, a sensible methodological imperative can be derived. It is not the consensus imperative, but something like it, though. In order to make full use of the social aspects, each scientist in a field of inquiry must have accurate beliefs about the beliefs of others in the field, and why they believe what they believe. If individual i believes X_i for reasons R_i, this leads to the first requirement that

C_1: $(\forall i, j \in \mathcal{P}) B_j B_i (X_i, R_i)$.

A rational scientists should make use of all available information, empirical and social, and should capitalise on the rationality of others. So a rational scientist j would form a new belief X_j^2 on the basis of all available information, or at least as much information as can be humanly processed. That is, X_j^2 should be formed on the basis of $\bigwedge_{i\in\mathcal{P}} B_i(X_i, R_i)$. The rationality of $B_j X_j^2$ thus requires at least $B_j \bigwedge_{i\in\mathcal{P}} B_i(X_i, R_i)$. Insofar as $B_j X_j^2$ is rational, this new belief can be informative to another scientists k. But k should be inclined to use $B_j(X_j^2)$ only if $B_k B_j [X_j^2 \wedge \bigwedge_{i\in\mathcal{P}} B_i(X_i, R_i)]$. So, the social process of using each-other's beliefs will function optimally when the requirement

C_2: $(\forall i, j, k \in \mathcal{P}) B_k B_j [X_j^2 \wedge B_i(X_i, R_i)]$

is met. The direction of the argument should be clear. At each stage, a scientist forms a belief on the basis of all beliefs in the group $\mathcal{P}$ at the previous stage. The scientist should be inclined to do so if these beliefs at the previous stage were formed on rational grounds, which at least requires that each group member believed what was believed in the group at the stage preceding the previous stage. So similarly to C_2, a requirement C_3 should be added, and C_4, etc. The optimality of the social process does not hinge on common belief, but something which might be called 'mutual understanding'. For a formal definition, denote the set of beliefs in $\mathcal{P}$ at stage k as

$$\mathcal{B}_{\mathcal{P}}^k \equiv \bigwedge_{i\in\mathcal{P}} B_i X_i^k.$$

Then $U_{\mathcal{P}}^k$, or mutual understanding at stage k, can be defined as

Definition 4 (Mutual understanding).
There is $U_{\mathcal{P}}^k$, or mutual understanding in a group $\mathcal{P}$ at stage k, if $\bigwedge_{i\in\mathcal{P}} B_i \bigwedge_{j=1}^{k-1} \mathcal{B}_{\mathcal{P}}^j$.

From my point of view, I should believe that you believe X_2^1 and that you believe that I believe X_1^1. In addition to this, I should later believe that you believe X_2^2 and that you believe I believe X_1^2. And in addition to this, I should later believe

that you believe X_2^3 and that you believe I believe X_1^3, and so forth. Certainly, this iterated process may lead to a convergence of your and my belief, and hence result in common belief. But this result is not required. The requirement is that beliefs and their reasons are public, and that everyone uses these public beliefs as they see fit.

6.6 Conclusion

Consensus and dissensus are natural outcomes when multiple persons focus on the same problem. That science is a social activity, in which individuals learn not only from empirical observations, but also from each other's interpretations of these observations, and evaluations of these interpretations, does not warrant the demand for consensus. Where people learn from each other, their beliefs will often show more similarity. But their beliefs do not have to be identical. Psychology's highly complex subject matter leaves plenty of room for difference in opinion. Of importance is how such differences are dealt with. Should they be smothered, by trying to find points where agreement is possible? Or, should they be stimulated, by trying to find points where disagreement is possible? The convergent strategy results in the collection of undisputed points of view, while the divergent strategy leads to the collection of disputable points of view. If the aim of the convergent strategy is not finding undisputed, but rationally indisputable points of view, the divergent strategy is a condition for the success of the convergent strategy. When it is realised that every belief is subject to improvement, and that the points of improvement can only be indicated by contradiction, with empirical observations or other beliefs, then the divergence in psychology is not her weakness, but rather her strength. Of course, where there is separation and dogmatism, where there are theoretical schools that isolate themselves from the rest of science, there the social process of science is impeded. But such disfunctioning can be attributed to an attempt to preserve an achieved consensus by denying the existence of a justified dissensus, rather than an inclination towards divergence as such.

Pluralism in belief and method is a result of the complexity of psychology's subject matter. To provide insight, theories need to be simplifications of the phenomena they describe. Reducing a complex problem to a manageable problem, focussing on certain aspects and ignoring others, can be done in different ways and result in different theories and methods. Sometimes, adherents to different methods or theories present their differences as a dilemma in which only one of the alternatives is right, while the different theories or methods are not truly contradictory, but instead focus on different aspects of a phenomenon. The examples of such quasi-dilemmas or schisms De Groot (1977) mentioned, such as correlational vs experimental research, and model-based (psychometric) measurement vs axiomatic (representational) measurement, are still relevant today. These are not contradictory methods in the sense that they usually lead to different results. One option is more stringent (experimental research and axiomatic measurement), but therefore also more limited in its application, than the other (correlational research and model-based measurement). While all research demands a choice from the possible approaches to the phenomenon under investigation, there is, as yet, no basis on which to lay down these choices for

all research focussing on the phenomenon. The application of different methods is not troublesome, but can only enrich the view of the subject of investigation. The consensus imperative, by which the hopeless search for one absolutism (justified true belief by compelling methodological rules) is replaced by another (justified true belief by unanimous consensus), is not the solution to psychology's complex problems.

Appendix

6A Observation regarding consensus as an epistemic criterion

In this appendix, the function of consensus as an epistemic criterion in dichotomous decision-problems is investigated. As in Chapter 5, we assume that a group of n individuals make individual decisions X_i in a dichotomous decision-problem, in which one of the alternatives, denoted as t, is objectively correct. The decision alternatives are valued -1 and 1. We assume that each individual in the group has a competence $p_i = P(X_i = t)$ in the problem, which can be decomposed as

$$p_i = \frac{e^{a(\theta_i - \beta)}}{1 + e^{a(\theta_i - \beta)}}, \tag{6A.1}$$

in which the person specific parameter θ_i reflects i's overall ability, and the problem specific parameter β the relative difficulty of the problem. Note that (6A.1) is identical to a Rasch or 1-parameter logistic model (Birnbaum, 1968) as encountered in item response theory (IRT). Assuming that the decisions are conditionally independent, so that

$$P(X_i = t \wedge X_j = t) = P(X_i = t)P(X_j = t)$$

for all i and j, the probability of consensus (in the sense of first-order unanimity, see Chapter 2) is

$$\prod_{i=1}^{n} p_i + \prod_{i=1}^{n} (1 - p_i).$$

This probability is decreasing in n, so it is highly unlikely that a large group of independently deciding individuals will be unanimous. Now we pose the following question: when does unanimity indicate that the group has arrived at the correct decision? To answer this question, assume that a group is unanimous in deciding $X_i = 1$ and consider the likelihood-ratio

$$LR = \frac{P(X_1 = 1, \ldots, X_n = 1 | t = 1)}{P(X_1 = 1, \ldots, X_n = 1 | t = -1)} = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n} (1 - p_i)} = e^{(a \sum_{i=1}^{n} \theta_i) - na\beta}.$$

Since $LR > 1$ indicates that it is more likely that the group arrived at the correct decision than that it arrived at the incorrect decision, it follows that consensus epistemologically informative only if $\beta < \overline{\theta}$. For relatively difficult problems, it is actually

more likely that the correct alternative is the opposite of the unanimous group decision! This is ironic, since it is generally for difficult problems that the requirement of consensus becomes more acute.

7
Summary and discussion

This thesis investigated the role of consensus in psychological methodology. This final chapter summarises the various issues raised and conclusions drawn in this investigation, and ends with some general thoughts on social learning and pluralistic methodology.

7.1 Consensus and methodology

Consensus has been a central concept in Western thinking on science. Classically, consensus was taken as a consequence of the scientific method, which demands unanimous consent amongst those who adopt it. Later theories in the philosophy of science, starting with Kuhn (1970), assigned a more pivotal role to consensus, arguing that consensus is itself arbiter in scientific decision problems. An important reason for raising the status of consensus is the general problem of underdetermination. It has been widely accepted that empirical evidence by itself is an insufficient basis to determine the choice between competing scientific theories. The classic problem of underdetermination concerns the logical possibility that, for any proposed theory to explain or describe regularities in a given body of data, there is an alternative, incompatible theory that is equally consistent with the given evidence. In Chapter 2, a more general version of underdetermination, called axiological underdetermination, was proposed. In contrast to certain beliefs that underdetermination is not a practical problem, since scientists base their decisions on more criteria than empirical adequacy, this thesis of underdetermination shows that it is not so easy to rid oneself of underdetermination. In the decision-theoretic framework introduced in Chapter 2, scientific inference is regarded as goal-directed behaviour. Scientists pursue epistemic aims, such as descriptive and predictive adequacy and simplicity, and the choice between competing

theories or methods is based on an evaluation of the utility of these theories and methods for those aspired aims. In such multi-attribute decision problems, underdetermination arises when multiple theories and methods have an identical utility. While axiological underdetermination is, as empirical underdetermination, a problem of possible existence, Chapter 3 showed how axiological underdetermination can actually arise in practical problems of statistical model selection. In a way, the addition of aims to empirical adequacy raises the likelihood of underdetermination, since there are multiple ways in which different values can be traded off against each other in multi-attribute decision problems, as opposed to single-attribute decision problems, where no such trade-off is possible. This chapter showed that methodological rules in general underdetermine theory choice. As such, the idea that theoretical consensus is a necessary consequence of methodological consensus should be abandoned.

Rather than being a direct consequence of the scientific method, some have argued that consensus is itself a basis for solving scientific decision problems. Since methodological rules are insufficient to determine theory choice, theorists such as Hesse (1980) have argued that social factors should be taken into account when explaining scientific decisions. A similar point is made in a major theory in social psychology. According to Festinger's (1950, 1954) theory of social comparison, when objective evidence is insufficient, individuals will attempt to validate their opinions by comparing them to those of others. From this theory, it follows that, if the choice between competing hypotheses is underdetermined by objective evidence, peer consensus on one of the alternative hypotheses should be a strong impetus to adopt this hypothesis. This prediction was tested in the two experiments described in Chapter 4. In an inductive rule discovery task, where the objective is to determine a logical rule that underlies a sequence of exemplars, the evidence was manipulated to result in different levels of underdetermination. In the first experiment it was shown that individuals tend to conform to an unanimously endorsed hypothesis if the rule is underdetermined by the evidence. In the second experiment, the rule was either strongly or weakly underdetermined by the given evidence, and it was expected that an increase in the level of underdetermination would result in more conformity to an unanimously held hypothesis. This prediction was supported. There was, however, an anomaly in the results of this experiment. For a weakly underdetermined rule, there were actually more individuals adopting a particular hypothesis when it was not unanimously endorsed than when it was unanimously endorsed. This may have been the result of a resistance to social pressure. While the hypothesis that greater levels of underdetermination result in a stronger tendency to conform to a unanimously endorsed position has been supported, this last anomaly indicates that consensus may not be directly aspired in matters where a single true belief is assumed to exist. Complementing this was the finding that in both experiments, the effect of the social information was small relative to that of the objective evidence. The assumption of a single true belief makes others' beliefs at once valuable and dispensable. Valuable, because all beliefs pertain to the same object and as such have potential informational value. Dispensable, because the truth of belief is not a function of the number of people sharing the belief. In situations of underdetermination, objective evidence is insufficient to delimit the number of plausible hypotheses to one. In such a situation, agreement with others may raise the plausibility of one of the underdetermined hypotheses. But,

if it is expected that underdetermination is only temporary – a consequence of the quality of present data and not the quality of all possible data – reliance on social validation will be temporary also. In the end, consensus is not an arbiter in scientific decision problems.

Chapter 5 investigated collective behaviour in nonmetric probability learning tasks. In general, there are two important reasons why a group can outperform individuals when it comes to making good judgements or decisions. The first is that individuals may possess (partly) non-overlapping information, so that the group as a whole can base its judgement or decision on more information than any individual alone. The second is that idiosyncratic biases may affect the group judgement or decision to a lesser extent than individual ones, because the idiosyncratic biases may cancel each other out in a group judgement or decision. While these two reasons render the assumption that groups are advantaged over individuals plausible, previous research has shown that groups often do not realise their potential. The two experiments of Chapter 5 were conducted in order to further scrutinise the assumption. In the first experiment, information was either completely distributed or shared, so that the effects of greater informational base and 'cancelling out' of individual bias could be separately investigated. As expected, groups in the distributed condition outperformed those in the shared condition. The effect of informational base was larger than that of cancelling-out individual bias. In the second experiment, information was partly shared and partly unique. Of concern in such situations is the so-called common knowledge effect. This refers to the finding that groups show a preference for shared information and tend to neglect unique information. A possible explanation of this effect, consistent with social comparison theory, is that the number of individuals sharing information is taken as an indication of the validity of the information. If this explanation is valid, the common knowledge effect should disappear if the shared information is known to have less validity than unique information. This hypothesis was confirmed, since there was no sign of the common knowledge effect in those conditions in which the shared information was of relatively low validity. Besides group achievement, the group process resulting in the collective decisions was of particular interest. Apart from those in the completely shared condition, groups appeared to adopt a weighting-by-confidence process, in which confidence in the correctness of responses determines the weight of those responses in the final collective response. Since confidence was related to conditional achievement (the probability of a correct response conditional on the given evidence), it was assumed that underlying this group process was a weighting-by-achievement process. This assumption did receive support, although the evidence was less marked than that for the weighting-by-confidence process.

Chapter 6 addressed the possible roles of consensus in a normative methodology. Three such roles were distinguished: consensus as a goal, as a means, and as a criterion. It was argued that consensus fulfills none of these roles adequately. Unqualified consensus is not a scientific goal, so it is necessary to distinguish between rational and other forms of consensus. However, this distinction is problematic. The proposed solution requires that at least one other goal is strived for besides consensus, which leads to the conclusion that consensus is at most a subordinate goal. As was further argued, consensus is an aspired goal only to the extent that other goals have already

been realised. When considering consensus as a means, one should be careful with one's claims. For instance, if consensus is taken as a means to arrive at true belief, the belief of a person adopting the consensus method can not itself be part of the consensus. If the person already concurs with the prevailing consensus, there is no effect of this consensus on that individual's belief. If the belief of the individual is discordant with the prevailing consensus, there was no overall consensus to begin with. Different effects of consensus have been mentioned in the literature. In general, a norm of critical dissensus has a better effect on the quality of collective judgements and decisions than a norm of consensus. When considering consensus as a scientific criterion, one can take consensus as a definition of truth, or a criterion of truth otherwise defined. The consensus theory of truth was shown to be highly problematic. When considering consensus as a criterion of truth otherwise defined, there is the problem that, in order to be valid, the criterion should actually not be applied. This is due to the fact that any influence the criterion has on the beliefs of those who adopt it, maps back to the criterion itself. Consensus as a criterion is, in this sense, self-referring. Consensus may be informative if it concerns agreement between independent sources. The belief of someone who conforms to the belief of someone else is redundant in this respect. Application of a consensus criterion, in such a way that individuals may change their belief accordingly, results in such redundancy. As such, the consensus criterion should be inconsequential – no one forms a belief on the basis of it – or not be applied. Consensus has no role in a normative methodology. Not as a goal, not as a method, not as a criterion.

7.2 Social learning and information integration

Both consensus and dissensus are natural states for a community of scientific inquirers. There is no proper rationale to demand either of them. While the consensus imperative is unfounded, it can not be denied that science is a social enterprise. But rather than focussing on the product of the social organisation of science – be it consensus or dissensus – it is more fruitful to focus on the process. What are the conditions under which the social process of scientific inquiry proceeds in an optimal fashion? And how should this social process ideally take form? Such questions have been largely neglected in theories of science. The philosophy of science, and epistemology in general, have almost exclusively focussed on the individual, while sociologists of science have, perhaps unwillingly, mainly addressed the 'irrational' aspects of science. A notable exception is the work of Merton, who took the social reward system of science as a primary factor in its success. As a psychologist, my interest is in more small scale social processes. How do scientists learn from each other? Is there a difference with everyday social learning? Such questions might be the topic of the 'social psychology of science', which is a recent – and as yet rather marginal – addition to science studies. Social psychology of science, as presented by Shadish and Fuller (1994), is an empirical study of how individual scientists are influenced by their social surroundings. The two empirical chapters of this thesis may be considered an addition to this programme.

While science studies offer descriptions of how scientific inferences proceeds, method-

ology is concerned with prescriptions for how it should proceed. Interpersonal learning should be a major topic of a social methodology of science. One reason to focus on social processes from a methodological viewpoint is the realisation that individual scientists, like everyone else, are subject to cognitive limitations. There are limits in memory and limits to the amount of information that can be adequately processed. Theoretical frameworks and methodological preferences will determine what, and how, new evidence is incorporated into an existing belief system. For these reasons, even the most enlightened scientist can only be expected to show a form of bounded rationality. But a community of such boundedly rational inquirers may, as a whole, possess a stronger form of rationality. If this is possible, it will surely depend on the social dynamics of the scientific enterprise. Chapter 5 showed that it may be advantageous for group members to focus on different information. If a phenomenon depends on a large variety of factors, the amount of relevant information may be overwhelming. In order to understand the phenomenon to some extent, the relevant information must be reduced to a manageable amount. Ignoring relevant evidence is not something to be proud of, as it will affect the validity of one's conclusions. Yet, by a division of cognitive labour, it may be possible to reach valid conclusions collectively, since a group as a whole may have focussed on all relevant evidence. Further research might indicate how such a division of cognitive labour can be fruitfully applied.

Others can fulfill a variety of functions in the social process of knowledge. Two important functions are as a source of information, and as a source of validation. Others are a source of information because it is impossible to acquire all knowledge first-hand from individual experience. One might consider consensus as a prerequisite for such information sharing, but it is not. Of course, information must be transmitted, and effective communication depends on some form of agreement between the conversing parties as to what a message means. But it is surely possible to understand a position without agreeing with it. How else would discussion and mutual criticism be possible? In psychology, inferences about unobservable psychological processes have to be based on observable behaviour. Often, the link between these is rather weak. The assumed psychological process is, if sufficient, usually not necessary for the occurrence of the observed behaviour. This is a fertile ground for problems of underdetermination. In one way, this is 'bad'. Underdetermination results in doubt, and doubt is uncomfortable. On the other hand, underdetermination can be quite useful, in the sense that proponents of different theories can make use of the same data. Underdetermination is a problem only insofar as a choice between the theories is necessary, and 'agreeing to disgree' forbidden. But if there is no demand for consensus, it is quite unproblematic that proponents of incompatible theories use the same data to support their incompatible positions.

Others can be a good source of validation, because it is easy to overlook problematic steps in one's reasoning. Consensus is surely not a prerequisite for others to function as a source of validation. Insofar as errors stem from bias in reasoning, or bias of methods, the less another shares a theoretical framework or methodology, the easier it will be for the other to detect those biases. Hence, this aspect of the social process of science is availed by dissensus.

When considering the conditions under which the social process of knowledge might optimally proceed, Chapter 6 argued that attempts should be made to reach

a situation of 'mutual understanding'. Mutual understanding was defined as a situation in which each individual in a group has accurate beliefs regarding the beliefs of other group members and regarding their reasons for holding those beliefs. In such a situation, it is possible to learn from others' beliefs on different levels. If all individuals in a group apply a social learning strategy, one can not only learn from others' initial beliefs, but also from others' later beliefs, which have been based on others' initial beliefs. To be more precise, a situation of mutual understanding can lead to the following type of reasoning:

> If alternative X is correct, I would expect group members $1, 2, \ldots, n$ to believe X with probability $p_1, p_2, \ldots, p_n$, respectively, for this is how I deem their competence in such matters. If $\neg X$ is correct, I would expect the same for this alternative. I have been informed of the beliefs of the group members and I deem these more likely given that X is the correct alternative than given that $\neg X$ is correct. So, based on this new information, I believe X. If my estimates of the group members' competence were correct, I would expect the others to now believe X as well. However, they now believe $\neg X$. So either their competence estimates are wrong, or mine. I deem it less likely that their competence estimates are all incorrect and mine correct than that their competence estimates are correct and mine wrong. Hence, I now believe $\neg X$ is more likely to be correct than X. So, I now believe $\neg X$.

A formalisation of this type of reasoning and its consequences has been worked out elsewhere[1], but the main thrust of the argument should be clear. In a situation of mutual understanding, one can make profitable use of others' reasoning in multiple ways. First of all, reasoning about others' knowledge can be a basis to form new or adjust existing beliefs. Second, if everyone applies such reasoning, one can use the resulting beliefs as a basis to validate this reasoning about others' knowledge. Now there are multiple levels at which idiosyncratic bias can be corrected for. Insofar as such bias is evenly distributed over all individuals in a group, there is a clear advantage of such a multi-layered process of social learning. This process will usually result in more uniformity of belief, but at no point is this a requirement. All that is asked is that each uses relevant information, both empirical and social, to form and adjust belief.

7.3 Prospects for a pluralistic methodology

This thesis began by noting psychology's uneasy status as a fragmented discipline, and a proposed remedy in the form of methodological unification. It should be clear by now that I do not endorse this solution. On the contrary, in the time I have spend thinking about consensus, I have become more and more a partisan of a pluralistic methodology.

An important argument for favouring such pluralism is based on the necessity of assumptions. As Coombs (1964) observed: 'we buy information with assumptions'(p.5).

[1] Speekenbrink, M. (to be submitted). Social validation for dichotomous decisions. A copy of this paper can be obtained by contacting the author.

There is no purely inductive, assumptionless science, no 'free lunch'. In order to make any inference, assumptions must be made which partly determine the outcome. Since there are no universally optimal methods, a rational choice for a particular method has to be based on a judgement regarding the nature of the phenomenon under investigation. As such, methods entail assumptions, which may or may not be directly testable. A rough but useful classification of assumptions can be based on the extent to which they are entrenched in a theory and the extent to which they are testable. Auxiliary hypotheses were already mentioned in the discussion of the Duhem-Quine thesis in Chapter 2. These are assumptions which are necessary to deduct predictions from a theory, but which do not directly belong to the theory itself. The assumption of a normal distribution was given as an example. A theory is not changed if the assumption of a normal distribution is replaced with the assumption of an exponential distribution, for example. If an assumption is auxiliary, one would like to generalise the conclusion to situations in which the particular distributional assumption is not made. One way of generalising inference over assumptions in the present example is by applying distribution-free, or nonparametric statistical techniques. Of course, such techniques are not devoid of assumptions, but the assumptions made are more general. If a nonparametric technique leads to the same conclusion as its parametric alternative, why not go for the more general option? I see no problem in testing a hypothesis twice, by different techniques, in order to investigate how assumptions pose limits on one's conclusions. Is the normality assumption necessary for reaching a particular conclusion, or not? And if pervasively so, should the assumption not be part of the theory, rather than considered auxiliary? This is not the place to discuss different types of assumptions and their role in scientific inference in more detail. The general point is that, if particular methods entail assumptions which do not belong the theory proper, one would like to generalise over assumptions. If particular methods entail assumptions, this is only possible by the application of multiple methods. Hence, there is good reason to adopt a pluralism of methodologies.

Pluralism in methodology is tied to the idea of triangulation. This term stems from navigation and refers to the fact that, in order to infer the precise position of an object in three-dimensional space, on needs accurate measurements of at least two other points in that space. In an analogue fashion, the thesis of methodological triangulation states that, in order to make valid inferences regarding a phenomenon under investigation, one needs to investigate the phenomenon with more than one method. Triangulation is now quite a mainstream concept in methodology, although it is rarely raised from the status of metaphor. An early exception is the Multitrait-Multimethod (MTMM) approach of Campbell and Fiske (1959) for assessing the validity of psychological measures. This approach has since its proposal received much attention, but its initial problems have not been solved (Fiske & Campbell, 1992). One problem to be dealt with is that different methods may lead to different conclusions. The notion of such 'method variance' was a main impetus for proposing the MTMM approach. How should one respond if different methods lead to a divergence in conclusions? There are a number of possible responses. One may conclude that one method leads to better results – because they are more consistent with one's expectations, for instance – and neglect the results of the other. A better response would be to look for the possible sources of the divergence. This exercise may result in more knowledge

regarding the methods, their underlying assumptions, and the phenomenon under study. If not neglected, anomalies are a major source of scientific progress.

7.4 Consensus, the last word?

When considering consensus as a goal of scientific inference, I have not yet mentioned one crucial distinction. From my point of view, I can aspire for you to agree with me, or I can aspire to agree with you. These are two quite different goals. If a consensus goal is held at all, it is probably in this first sense. Can one found a methodology on the first principle 'agree with me'? I think not. Yet, such a principle may be quite basic to science. Scientists are not agnostic when they start a research project. Most have a firm belief in a theoretical position. As such, they need no empirical evidence to persuade them of the correctness of their position. Empirical evidence may mainly serve the purpose of persuading others. If consensus is a scientific goal in this respect, I have no reason to argue against it, for it will only lead to more discussion and mutual criticism.

Consensus is a strange animal. I have put forward the thesis that we should not strive for scientific consensus. But in doing so, I hope that my arguments are convincing enough to persuade you to agree. Yet, to remain truthful to my position, I can only hope you have good reasons to disagree.

References

Agassi, J. (1975). *Science in flux*. Dordrecht: D. Reidel.

Akaike, H. (1992[1973]). Information theory and an extension of the maximum likelihood principle. In S. Kotz & K. L. Johnson (Eds.), *Breakthroughs in statistics. Vol 1* (pp. 610–624). London: Springer-Verlag.

Allen, V. L. (1965). Situational factors in conformity. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. II, pp. 133–175). New York: Academic Press.

Allen, V. L., & Levine, J. M. (1971). Social support and conformity: The role of independent assessment of reality. *Journal of Experimental Social Psychology*, *7*, 48–58.

Allen, V. L., & Wilder, D. A. (1977). Social comparison, self-evaluation, and conformity to the group. In J. Suls & R. L. Miller (Eds.), *Social comparison processes: Theoretical and empirical perspectives* (pp. 187–208). Washington: Hemisphere.

Andersson, H., & Brehmer, B. (1979). Note on the policies acquired in interpersonal learning. *Organizational Behavior and Human Performance*, *24*, 195–201.

Asch, S. E. (1952). *Social psychology*. New York: Prentice Hall.

Attneave, F. (1959). *Applications of information theory to psychology*. New York: Holt, Rinehart and Winston.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, *4*, 1236–1239.

Bach, K. (1975). Analytic social philosophy – basic concepts. *Journal for the Theory of Social Behaviour*, *5*, 189–214.

Baigrie, B. S., & Hattiangadi, J. N. (1992). On consensus and stability in science. *British Journal for the Philosophy of Science*, *43*, 435–458.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.

Berend, D., & Paroush, J. (1998). When is Condorcet's jury theorem valid? *Social Choice and Welfare*, *15*, 481–488.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading: Addison-Wesley.

Bjerring, A. K., & Hooker, C. A. (1980). Lehrer, consensus and science: The empiricist watershed. In R. J. Bogdan (Ed.), *Keith Lehrer* (pp. 181–203). Dordrecht: D. Reidel.

Björkman, M. (1973). Inference behavior in nonmetric ecologies. In L. Rappoport &

D. A. Summers (Eds.), *Human judgement and social interaction* (pp. 144–168). New York: Holt, Rinehart and Winston.

Bliese, P. D., & Halverson, R. R. (1998). Group consensus and psychological well-being: A large field study. *Journal of Applied Social Psychology*, *28*(7), 563–580.

Boland, P. J. (1989). Majority systems and the Condorcet jury theorem. *The Statistician*, *38*, 181–189.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory and Psychology*, *14*, 105–120.

Bower, J. L. (1965). Group decision making: A report of an experimental study. *Behavioral Science*, *10*, 277–289.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, *44*, 62–91.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*, 260–273.

Brehmer, B. (1973). Effects of task predictability and cue validity on interpersonal learning of inference tasks involving both linear and nonlinear relations. *Organizational Behavior and Human Performance*, *10*, 24–46.

Brehmer, B. (1974). The effect of cue intercorrelation on interpersonal learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *12*, 397–412.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217.

Bunge, M. (1962). The complexity of simplicity. *The Journal of Philosophy*, *59*, 113–135.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Campbell, N. (1921). *What is science?* London: Methuen.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Castellan, N. J. (1974). The effect of different types of feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *11*, 44–64.

Castellan, N. J. (1977). Decision making with multiple probabilistic cues. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory, vol. 2* (pp. 117–147). Hillsdale: Erlbaum.

Cheeseman, P. (1990). On finding the most probable model. In J. Schrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 73–95). San Mateo: Morgan Kaufman.

Chomsky, N. (1957). *Syntactic structures*. Den Haag: Mouton.

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences*, *21*, 169–239.

Clark, H. H., & Marshall, C. R. (1981). Defnite reference and mutual knowledge. In A. K. Joshi (Ed.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190.

Cooksey, R. W. (1996). *Judgment analysis.* San Diego: Academic Press.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cutting, J. E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology*, *44*, 3–19.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* New York: Cambridge University Press.

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, *80*, 97–125.

Denmark, F. L., & Krauss, H. H. (2005). Unification through diversity. In R. J. Sternberg (Ed.), *Unity in psychology: Possibility or pipedream* (pp. 15–36). Washington: American Psychological Association.

Derkse, W. (1993). *On simplicity and elegance.* Delft: Eburon.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, *51*, 629–636.

Dobson, A. J. (2002). *An introduction to generalised linear models* (2nd ed.). Boca Raton: Chapman & Hall.

Earle, W. B. (1986). The social context of social comparison: Reality versus reassurance? *Personality and Social Psychology Bulletin*, *12*, 159–168.

Edgell, S. E. (1978). Configural information processing in two-cue probability learning. *Organizational Behavior and Human Performance*, *22*, 404–416.

Edgell, S. E. (1980). Higher order configural information processing in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Performance*, *25*, 1–14.

Edgell, S. E. (1993). Using configural and dimensional information. In N. J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 43–64). Hillsdale: Lawrence Erlbaum.

Ellis, B. (1988). Solving the problem of induction using a values-based epistemology. *British Journal for the Philosophy of Science*, *39*, 141–160.

Eysenck, H. J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and Social Psychology*, *73*, 1224–1237.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge.* Cambridge: MIT press.

Festinger, L. (1950). Informal social communication. *Psychological Review*, *57*, 271–281.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–140.

Feyerabend, P. (1974). Consolidations for the specialist. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 197–230). London: Cambridge University Press.

Feyerabend, P. (1975). *Against method.* London: Verso.

Fine, A. (1996). Science made up: Constructivist sociology of scientific knowledge.

In P. Galison & D. J. Stump (Eds.), *The disunity of science* (pp. 231–254). Stanford: Stanford University Press.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, *112*, 393–395.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205–231.

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, *45*, 1–35.

Fraassen, B. C. van. (1980). *The scientific image*. Oxford: Oxford University Press.

Fraley, C., & Raftery, A. E. (2002). *MCLUST: Software for model-based clustering, discriminant analysis and density estimation* (Tech. Rep. No. 415R). Washington, United States of America: Department of Statistics, University of Washington.

Geanakoplos, J. (1992). Common knowledge. *Journal of Economic Perspectives*, *6*, 53–82.

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist*, *40*(3), 266–275.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chigaco Press.

Gigone, D., & Hastie, R. (1993). The common knowledge effect: Information sharing and group judgement. *Journal of Personality and Social Psychology*, *65*, 959–974.

Goethals, G. R. (1972). Consensus and modality in the attribution process: The role of similarity and information. *Journal of Personality and Social Psychology*, *21*, 84–92.

Goethals, G. R., & Darley, J. (1977). Social comparison theory: An attributional approach. In J. Suls & R. L. Miller (Eds.), *Social comparison processes: Theoretical and empirical perspectives* (pp. 259–278). Washington: Hemisphere.

Goethals, G. R., Darley, J. M., & Kriss, M. (1978). The impact of opinion agreement as a function of the ground for agreement. *Representative Research in Social Psychology*, *9*, 30–42.

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, *44*, 153–170.

Goodman, N. (1954). *Fact, fiction and forecast*. London: Athlone Press.

Gorenflo, D. W., & Crano, W. D. (1989). Judgemental subjectivity/objectivity and locus of choice in social comparison. *Journal of Personality and Social Psychology*, *57*, 605–614.

Groot, A. D. de. (1961). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen* (3 ed.). Den Haag: Mouton.

Groot, A. D. de. (1971). *Een minimale methodologie op sociaal-wetenschappelijke basis*. Den Haag: Mouton.

Groot, A. D. de. (1977). Gevraagd: Forum-convergentie inzake begrips- theorie- en besluitvorming. *Nederlands Tijdschrift voor de Psychologie*, *32*(4), 219–241.

Groot, A. D. de. (1982). *Academie en forum: Over hoger onderwijs en wetenschap*. Meppel: Boom.

Groot, A. D. de. (1990a). Unifying psychology: A European view. *New Ideas in Psychology*, *8*, 309–320.

Groot, A. D. de. (1990b). Unifying psychology: Its preconditions. In W. J. Baker, M. E. Hyland, R. van Hezewijk, & S. Terwee (Eds.), *Recent trends in theoretical psychology. Volume II* (pp. 1–36). New York: Springer-Verlag.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

Gustin, B. H. (1973). Charisma, recognition, and the motivation of scientists. *American Journal of Sociology*, *78*, 1118–1134.

Hacking, I. (1999). *The social construction of what?* Cambridge: Harvard University Press.

Hammond, K. R., Wilkins, M. M., & Todd, F. J. (1966). A research paradigm for the study of interpersonal learning. *Psychological Bulletin*, *65*, 221–232.

Hedeker, D., & Gibbons, R. D. (1996). Mixor: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157–176.

Heerden, J. van. (1980). De overbodige strijd om unanimiteit. *Tijdschrift voor Onderwijsresearch*, *5*(extra nummer), 35–39.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: The Free Press.

Hesse, M. (1980). *Revolutions and reconstructions in the philosophy of science*. Brighton: The Harvester Press.

Hibberd, F. J. (2001). Gergen's social constructivism, logical positivism and the continuity of error. Part 1: Conventionalism. *Theory & Psychology*, *11*, 297–321.

Hill, G. W. (1982). Group versus individual performance: Are n + 1 heads better than one? *Psychological Bulletin*, *91*(3), 517–539.

Holloman, C. R., & Hendrick, H. W. (1972). Adequacy of group decisions as a function of the decision-making process. *Academy of Management Journal*, *15*(2), 175–184.

Horowitz, I. L. (1962). Consensus, conflict and cooperation: A sociological inventory. *Social Forces*, *41*, 177–188.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hume, D. (1748/1910). *An enquiry concerning human understanding*. New York: P.F. Collier & Son.

Insko, C. A., Drenan, S., Solomon, M. R., Smith, R., & Wade, T. J. (1983). Conformity as a function of the consistency of positive self-evaluation with being liked and being right. *Journal of Experimental Social Psychology*, *19*, 341–358.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Oxford: Houghton Mifflin.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

Jones, E. E., & Gerard, H. B. (1967). *Foundations of social psychology*. New York: Wiley.

Kalberer, J. T. (1985). Peer review and the consensus development process. *Science, Technology, and Human Values*, *10*, 63–72.

Kanazawa, S. (1998). A brief note on a further refinement of the Condorcet jury theorem for heterogeneous groups. *Mathematical Social Sciences*, *35*, 69–73.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks: Sage.

Katzenstein, G. (1996). The debate on structured debate: Toward a unified theory. *Organizational Behavior and Human Decision Processes*, *66*, 316–332.

Keeny, R. L., & Raiffa, H. (1976). *Decisions with multiple objective: Preferences and value tradeoffs.* New York: Wiley.

Kendler, H. H. (2002). Romantic versus realistic views of psychology. *American Psychologist*, *57*, 1125–1126.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgement: Comparing individuals and groups. *Psychological Review*, *103*, 687–719.

Kerr, N. L., Niedermeier, K. E., & Kaplan, M. F. (1999). Bias in jurors vs bias in juries: New evidence from the SDS perspective. *Organizational Behavior and Human Decision Processes*, *80*, 70—86.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions.* New York: Oxford University Press.

Klazinga, N. S., Everdingen, J. J. E. V., & Casparie, A. F. (1989). Klinische epidemiologie en consensus. *Tijdschrift voor Sociale Gezondheidszorg*, *67*, 340–342.

Knorr-Cetina, K. (1995). Laboratory studies: The cultural approach to the study of science. In S. Jasanoff, G. E. Markle, J. C. Petersen, & T. Pinch (Eds.), *Handbook of science and technology studies* (pp. 140–166). Thousand Oaks: Sage.

Koch, S. (1981). The nature and limits of psychological knowledge: Lessons from a century qua "science". *American Psychologist*, *36*, 257–269.

Koch, T., & Ridgley, M. (2000). The Condorcet's jury theorem in a bioethical context: The dynamics of group decision making. *Group Decision and Negotiation*, *9*, 379–392.

Krantz, D. L. (1987). Psychology's search for unity. *New Ideas in Psychology*, *5*, 329–339.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Kuhn, T. (1977). *The essential tension.* Chicago: University of Chicago Press.

Kukla, A. (2000). *Social constructivism and the philosophy of science.* London: Routledge.

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, *36*, 617–634.

Larson, J. R., Christensen, C., Abbott, A. S., & Franz, T. M. (1996). Diagnosing groups: Charting the flow of information in medical decision-making teams. *Journal of Personality and Social Psychology*, *71*, 315–330.

Laudan, L. (1984). *Science and values.* Berkeley: University of California Press.

Laudan, L. (1996). *Beyond positivism and relativism: Theory, method, and evidence.* Boulder: Westview Press.

Laughlin, P. R. (1999). Collective induction: Twelve postulates. *Organisational Behavior and Human Decision Processes*, *80*, 50–69.

Laughlin, P. R., & Bonner, B. L. (1999). Collective induction: Effects of multiple hypotheses and multiple evidence in two problem domains. *Journal of Personality and Social Psychology*, *77*, 1163–1172.

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, *88*, 605–620.

Laughlin, P. R., Chandler, J. S., Shupe, E. J., Magley, V. J., & Hulbert, L. G. (1995). Generality of a theory of collective induction: Face-to-face and computer-mediated interaction, amount of potential information, and group versus member choice of evidence. *Organizational Behavior and Human Decision Processes*, *63*, 98–111.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177–189.

Leeuw, J. de. (1990). Data modeling and theory construction. In J. Hox & A. de Jonge-Gierveld (Eds.), *Operationalization and research strategy* (p. 229-246). Amsterdam: Swetz & Zeitlinger.

Leeuw, J. de. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In S. Kotz & K. L. Johnson (Eds.), *Breakthroughs in statistics vol. 1* (pp. 599–609). London: Springer-Verlag.

Levi, I. (1967). *Gambling with truth.* London: Routledge & Kegan Paul.

Levi, I. (1980). *The enterprise of knowledge.* Cambridge: MIT Press.

Levine, J. M., Resnick, L. B., & Higgins, E. T. (1993). Social foundations of cognition. *Annual Review of Psychology*, *44*, 585–612.

Levine, J. M., & Thompson, L. (1996). Conflict in groups. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 745–776). New York: Guilford.

Lewis, D. K. (1969). *Convention: A philosophical study.* Cambridge: Harvard University Press.

Li, M., & Vitáni, P. M. B. (1992). Inductive reasoning and Kolmogorov complexity. *Journal of Computer and System Sciences*, *44*, 343–384.

Liebrucks, A. (2001). The concept of social construction. *Theory & Psychology*, *11*, 363–391.

Linhart, H., & Zucchini, W. (1986). *Model selection.* New York: John Wiley.

Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. *Psychometrika*, *20*, 139-148.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*, 185–199.

Markus, K. A. (2002). Statistical equivalence, semantic equivalence, eliminative

induction, and the Raykov-Marcoulides proof of infinite equivalence. *Structural Equation Modeling*, *9*, 503–522.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a mutlivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*, 247–255.

McGarty, C., Turner, J. C., Oakes, P. J., & Haslam, S. A. (1993). The creation of uncertainty in the influence process: The roles of stimulus information and disagreement with similar others. *European Journal of Social Psychology*, *23*, 17–38.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, *19*, 97–116.

Meehl, P. E. (1999). How to weight scientists' probabilities is not a big problem: Comment on Barnes. *British Journal for the Philosophy of Science*, *50*, 283–295.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*, 639–667.

Miller, C. E. (1989). The social psychological effects of group decision rules. In P. B. Paulus (Ed.), *Psychology of group influence* (2nd ed., pp. 327–355). Hillsdale: Lawrence Erlbaum Associates.

Molenaar, P. C. M., & Von Eye, A. (1994). On the arbitrary nature of latent variables. In A. Von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 226–242). Thousand Oaks: Sage.

Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, *12*(1), 3–23.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.

Nagel, E., & Newman, J. R. (1958). *Gödel's proof.* New York: New York University Press.

Nelson, A. (1994). How could scientific facts be socially constructed? *Studies in History and Philosophy of Science*, *25*, 535–547.

Nerlich, B. (2004). Coming full (hermeneutic) circle: The controversy about psychological methods. In Z. Todd, B. Nerlich, S. McKeown, & D. D. Clarke (Eds.), *Mixing methods in psychology* (pp. 17–36). Hove: Psychology Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nicolis, G., & Prigogine, I. (1989). *Exploring complexity.* New York: Freeman.

Nitzan, S., & Paroush, J. (1982). Optimimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, *23*, 289–297.

Olson, J. M., Ellis, R. J., & Zanna, M. P. (1983). Validating objective versus subjective judgements: Interest in social comparison and consistency information. *Personality and Social Psychology Bulletin*, *9*, 427–436.

Orive, R. (1988). Social projection and social comparison of opinions. *Journal of Personality and Social Psychology*, *54*, 953–964.

Owen, G., Grofman, B., & Feld, S. L. (1989). Proving a distribution-free generalisation of the Condorcet jury theorem. *Mathematical Social Sciences*, *17*, 1–16.

Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly*, *12*, 286–302.

Poincaré, H. (1905/1979). *Wetenschap en hypothese*. Meppel: Boom.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Routledge.

Popper, K. R. (1974). Normal science and its dangers. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 51–58). London: Cambridge University Press.

Postmes, T., Spears, R., & Cihangir, S. (2001). Quality of decision making and group norms. *Journal of Personality and Social Psychology*, *80*, 918–930.

Quine, W. V. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*, 20–43.

Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis*, *9*, 313–328.

Raykov, T., & Penev, S. (1999). On structural model equivalence. *Multivariate Behavior Research*, *34*, 199–244.

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge*. Chicago: University of Chicago Press.

Rescher, N. (1993). *Pluralism: Against the demand for consensus*. Oxford: Clarendon Press.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, *11*, 416–431.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE transactions on information theory*, *42*, 40–47.

Royce, J. R. (1970). *Toward unification in psychology: The first Banff conference on theoretical psychology*. University of Toronto Press: Toronto.

Royce, J. R. (1987). More order than a telephone book? A response to Krantz. *New Ideas in Psychology*, *5*, 341–345.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *The American Economic Review*, *79*, 385–391.

Rychlak, J. F. (2005). Unification in theory and method: Possibilities and impossibilities. In R. J. Sternberg (Ed.), *Unity in psychology: Possibility or pipedream?* (pp. 145–157). Washington: American Psychological Association.

Scheff, T. J. (1967). Toward a sociological model of consensus. *American Sociological Review*, *32*, 32–46.

Schiffer, S. R. (1972). *Meaning*. Oxford: Oxford University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*, 461–464.

Shadish, W. R., & Fuller, S. (1994). *The social psychology of science*. New York: The Guilford Press.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 397–423,623–656.

Sober, E. (1975). *Simplicity*. Oxford: Clarendon Press.

Sober, E. (2000). Simplicity. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 433–441). Oxford: Blackwell.

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group performance depends on the majority rule. *Psychological Science*, *9*, 456–463.

Staats, A. W. (1991). Unified positivism and unification psychology. *American Psychologist*, *46*, 899–912.

Staats, A. W. (1999). Unifying psychology requires new infrastructure, theory, method, and a research agenda. *Review of General Psychology*, *3*, 3–13.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*, 1467–1478.

Stasson, M. F., Kameda, T., Parks, C. D., & Zimmerman, S. K. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly*, *54*, 25–35.

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*, 309–331.

Sternberg, R. J., & Grigorenko, E. L. (2001). Unified psychology. *American Psychologist*, *56*, 1069–1079.

Suls, J., Martin, R., & Wheeler, L. (2000). Three kinds of opinion comparison: The triadic model. *Personality and Social Psychology Review*, *4*, 219–237.

Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, *4*, 341–376.

Thorburn, W. M. (1918). The myth of Occam's razor. *Mind*, *27*, 345–353.

Tindale, R. S. (1993). Decision errors made by individuals and groups. In N. J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 109–124). Hillsdale: Lawrence Erlbaum.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, *17*, 528–530.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Williams, L. J., Bozdogan, H., & Aiman-Smith, L. (1996). Inference problems with equivalent models. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 279–314). Mahwah: Lawrence Erlbaum Associates.

Wittenbaum, G. M., & Park, E. S. (2001). The collective preference for shared information. *Current Directions in Psychological Science*, *10*(2), 70–73.

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123–136.

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology*, *33*, 345–366.

Ziman, J. (1968). *Public knowledge: An essay concerning the social dimension of science*. New York: Cambridge University Press.

Ziman, J. (1978). *Reliable knowledge: An exploration of the grounds for belief in science*. Cambridge: Cambridge University Press.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*, 41–61.

Consensus en methodologie
Nederlandse samenvatting

Consensus is een centraal begrip in Westerse ideeën over de wetenschap. Vroeger werd verondersteld dat consensus een direct gevolg is van de Wetenschappelijke Methode, die unanieme instemming vereist onder diegenen die haar toepassen. Latere wetenschapsfilosofische theorieën, zoals die van Kuhn (1970), kennen consensus een meer fundamentele rol toe. In plaats van gevolg, is consensus hier een oorzaak in het proces van wetenschappelijke inferentie. Een belangrijke reden om dit standpunt in te nemen is het probleem van onderdeterminatie. Het is nu een algemeen geaccepteerd gegeven dat empirisch bewijs alleen niet voldoende is om de keuze tussen wetenschappelijke theorieën te bepalen. In de klassieke vorm betreft onderdeterminatie de logische mogelijkheid dat er, voor een willekeurige theorie die consistent is met empirisch bewijs, een alternatieve theorie bestaat die even consistent is met dat bewijs. Als zodanig kan een specifieke keuze voor een theorie niet worden gerechtvaardigd. Sommige auteurs stellen dat onderdeterminatie geen praktisch probleem is; wetenschappers houden er immers andere doelen op na dan empirische adequaatheid alleen. Als antwoord op deze stelling wordt in hoofdstuk 2 een meer algemene versie van onderdeterminatie gegeven, die axiologische onderdeterminatie wordt genoemd. In het besliskundig raamwerk dat in dit hoofdstuk wordt geïntroduceerd wordt wetenschappelijk handelen als doelgericht handelen opgevat. Wetenschappers streven cognitieve doelen na, zoals descriptieve en predictieve adequaatheid, simpelheid, e.d., en de keuze tussen wetenschappelijke theorieën en methoden berust op de evaluatie van hun prestatie met betrekking tot deze doelen. In zulke beslisproblemen met meerdere attributen treedt onderdeterminatie op wanneer verschillende theorieën of methoden een identieke utiliteit hebben. Hoewel deze axiologische onderdeterminatie, net als empirische onderdeterminatie, allereerst een probleem van logische existentie betreft, wordt in hoofdstuk 3 aangetoond hoe deze vorm van onderdeterminatie ook in praktische gevallen kan voorkomen.

In hoofdstuk 3 wordt het probleem van statistische modelselectie geanalyseerd aan de hand van een theorie over de structuur van wetenschappelijke discussies. Deze theorie, de hiërarchische theorie van rechtvaardiging, stelt dat wetenschappelijke onenigheid op meerdere niveaus kan worden beëindigd. De theorie stelt dat onenigheid over feitelijke zaken – opgevat in een ruime zin van stellingen, hypothesen, theorieën en modellen aangaande fenomenen in de werkelijkheid – kan worden opgelost door consensus te bereiken op methodologisch niveau. Onenigheid op methodologisch vlak kan op haar beurt worden opgelost door consensus te bereiken op axiologisch niveau, het niveau van wetenschappelijke doelen. De hiërarchische theorie sluit goed aan bij het besliskundig raamwerk van hoofdstuk 2, maar hoewel de theorie een inzichtelijke beschrijving geeft van aspecten van wetenschappelijke inferentie en discussie, blijkt het geen adequate normatieve theorie. De scheiding tussen theorie en methode is veelal niet zo sterk te trekken, aangezien de optimaliteit van een specifieke methode afhangt van eigenschappen van het fenomeen dat ermee wordt bestudeerd. De keuze voor een methode dient dan ook te berusten op theoretische ideeën of assumpties betreffende het fenomeen en niet alleen op aangehouden axiologische doelen. Daarnaast speelt onderdeterminatie de theorie parten. Ook al is consensus bereikt over de juiste selectiemethode, meerdere tegenstrijdige modellen kunnen als even goed worden beoordeeld op het criterium van deze methode. Als zodanig dient de stelling dat theoretische consensus een direct gevolg van methodologische consensus te worden verworpen. Dit hoofdstuk laat zien dat theorieën niet alleen worden ondergedetermineerd door empirische data, en daarmee door methoden waarmee de consistentie van een theorie met observaties wordt bepaald, maar dat methoden tevens ondergedetermineerd zijn door axiologische doelen.

Sommigen, zoals Hesse (1980) en Kuhn (1970, 1977), stellen dat consensus geen direct gevolg van de wetenschappelijke methode is, maar juist een basis vormt voor de oplossing van wetenschappelijke beslisproblemen. Aangezien methodologische regels geen voldoende grond geven om wetenschappelijke keuzen te rechtvaardigen, dienen sociale factoren in acht te worden genomen om deze keuzen te verklaren. Een overeenkomstig idee is verwoord in een belangrijke theorie uit de sociale psychologie. Volgens Festinger's (1950, 1954) sociale vergelijkingstheorie trachten mensen, indien de objectieve werkelijkheid hiervoor ontoereikend is, hun ideeën en opinies te valideren door ze met die van anderen te vergelijken. Vanuit deze theorie kan worden afgeleid dat, indien hypothesen ondergedetermineerd zijn, unanimiteit tussen andere individuen betreffende de juiste hypothese een sterke stimulans zal zijn om deze hypothese over te nemen. Deze verwachting wordt getoetst in de twee experimenten die in hoofdstuk 4 worden besproken. In inductieve regel-ontdekkingstaken is het doel te achterhalen welke logische regel ten grondslag ligt aan een sequentie van observaties. In zulke taken kan de mate waarin de regel ondergedetermineerd is worden gemanipuleerd door middel van kleine verschillen in de aangeboden sequenties. In het eerste experiment is de regel ondergedetermineerd of gedetermineerd, en is sprake van consensus tussen de andere proefpersonen betreffende de correcte hypothese of juist onenigheid. De verwachting dat proefpersonen zich conformeren aan een consensuele hypothese indien de regel ondergedetermineerd is werd geconfirmeerd. Voor gedetermineerde regels werd nauwelijks evidentie voor zulk conformisme gevonden, zodat kan worden gesteld dat de hypothesen van anderen voornamelijk een informationele, en niet zozeer een

normatieve, invloed hebben. In het tweede experiment waren de regels in alle gevallen ondergedetermineerd, maar in verschillende mate. Er werd ondersteuning gevonden voor de hypothese dat de neiging tot conformeren aan een consensuele hypothese sterker is naarmate de onderdeterminatie van de regel sterker is. Onverwacht was echter dat, voor in lichte mate ondergedetermineerde regels, meer mensen voor een specifieke hypothese kozen wanneer deze geen unanieme instemming van de andere proefpersonen kreeg. Dit kan duiden op een mogelijk verzet tegen sociale beïnvloeding. In ieder geval kan niet zonder meer worden aangenomen dat het vormen van consensus een direct doel is in situaties waarin een objectief correct antwoord wordt geacht te bestaan. De aanname van een enkele correcte oplossing leidt er toe dat de ideeën van anderen tegelijkertijd waardevol en ontbeerlijk zijn. Ze zijn waardevol omdat een ieders ideeën betrekking hebben op hetzelfde object of doel. Ze zijn ontbeerlijk omdat de waarheid van een hypothese niet afhangt van het aantal mensen dat de hypothese ondersteunt. In beide experimenten was het effect van de sociale informatie relatief klein in vergelijking met het effect van de empirische informatie. In situaties van onderdeterminatie is empirisch bewijs onvoldoende om het aantal plausible hypothesen te beperken tot één. In zulke situaties kan instemming van anderen met een hypothese de plausibiliteit van deze hypothese versterken. Maar indien verwacht wordt dat de onderdeterminatie tijdelijk is – een gevolg van de kwaliteit van de huidige data, en niet van de kwaliteit van alle mogelijke data – dan zal de berusting op sociale factoren ook van tijdelijke aard zijn. De uiteindelijke scheidsrechter is dan toch de empirie, en niet consensus.

In hoofdstuk 5 wordt collectief gedrag in zogenaamde 'nonmetric multiple cue probability learning' (NMCPL) taken onderzocht. Dit zijn taken waarin, aan de hand van observaties van nominale variabelen, een nominaal criterium moet worden voorspeld. In het algemeen zijn er twee belangrijke redenen waarom groepen in zulke situaties tot betere beslissingen kunnen komen dan individuen. De eerste is dat individuen over (gedeeltelijk) verschillende informatie beschikken, zodat de groep als geheel een beslissing op meer informatie kan baseren dan ieder individu alleen. De tweede reden is dat idiografische vertekeningen in het oordeels- of beslissingsproces een grotere invloed op individuele dan op collectieve beslissingen kunnen uitoefenen. Het eerste experiment was zo opgesteld dat de validiteit van deze twee assumpties afzonderlijk kon worden onderzocht. In een conditie beschikte ieder groepslid over unieke informatie en in de andere conditie beschikte ieder groepslid over dezelfde informatie. Zoals verwacht presteerden groepen in de eerste conditie beter dan in de tweede. Als zodanig kan worden gesteld dat het effect van de rijkere informatie-basis op de kwaliteit van groepsbeslissingen groter is dan het effect van het uitmiddelen van idiosyncratische vertekeningen. In het tweede experiment was de informatie gedeeltelijk uniek en gedeeltelijk gedeeld. In zulke situaties is veelal sprake van het zogenaamde 'common knowledge' effect, hetgeen betekent dat groepen zich voornamelijk richten op gedeelde informatie en unieke informatie negeren. Een mogelijke verklaring voor dit effect, consistent met de sociale vergelijkingstheorie, is dat het aantal personen dat over bepaalde informatie beschikt wordt opgevat als indicatie van de validiteit van die informatie. Hieruit kan worden afgeleid dat, indien andere indicaties van de validiteit van informatie beschikbaar zijn, het 'common knowledge' effect niet zal optreden. Deze hypothese werd ondersteund: de gedeelde informatie had een grote

invloed op de collectieve beslissing indien deze een relatief hoge validiteit had, maar niet indien de validiteit van de gedeelde informatie relatief laag was. Behalve de evaluatie van de prestatie van groepen in NMCPL taken richtte het onderzoek zich op het groepsproces dat resulteert in collectieve beslissingen. Afgezien van groepen in de conditie waarin informatie geheel gedeeld was, was het groepsproces een vorm van 'wegen-door-zekerheid', waar het gewicht van individuele beslissingen in de collectieve beslissing een functie is van de subjectieve zekerheid dat de individuele beslissingen correct zijn. Aangezien de subjective zekerheid gerelateerd was aan de eerdere prestatie van de individuen bij de hen aangeboden informatie, lijkt onderliggend aan dit groepsproces een vorm van 'wegen-naar-prestatie' te liggen. Voor dit laatste proces werd ondersteuning gevonden, alhoewel het bewijs minder sterk was dan voor het wegen-naar-zekerheid proces.

De rol van consensus in een normatieve methodologie wordt kritisch beschouwd in hoofdstuk 6. Drie mogelijke rollen worden onderscheiden: consensus als doel, als middel en als criterium. Er wordt beargumenteerd dat consensus geen van deze rollen vervult. Als doel betreft consensus niet elke consensus. Het is noodzakelijk rationale consensus van andere vormen van consensus te onderscheiden, maar dit onderscheid is problematisch. De voorgestelde oplossing vereist dat naast consensus tenminste een ander doel wordt nagestreefd. Er wordt gesteld dat indien consensus een wetenschappelijk doel is, het alleen wordt nagestreefd indien andere doelen zijn bereikt. Zo is consensus ten hoogste een ondergeschikt doel. Wordt consensus als middel opgevat, dan is enige voorzichtigheid geboden. Is consensus een middel, dan dient het iets te bewerkstelligen. Beschouwt men consensus bijvoorbeeld als middel voor het bereiken van ware opinies, dan mogen de opinies van diegenen die het middel toepassen geen onderdeel van de consensus zijn. Indien de persoon al met de consensus instemt, is er immers geen effect van consensus. Indien de persoon niet met de consensus instemt, was er eigenlijk geen sprake van consensus. Consensus kan zo bezien alleen middel zijn voor iemand zonder noemenswaardige opinie aangaande het onderwerp van consensus. In de psychologische literatuur worden verschillende effecten van consensus genoemd. Uit onderzoek blijkt dat een norm van kritische discussie over het algemeen beter is dan een norm van consensus. Daarmee is de consensus-imperatief een minder goed middel dan andere imperatieven, zoals rationele dissensus en kritische discussie. Wordt consensus als criterium gehanteerd, dan zijn er in ieder geval twee mogelijkheden: consensus kan als waarheidsdefinitie worden gebruikt, of als indicatie van een op andere wijze gedefinieerde waarheid. De consensustheorie van de waarheid kent ernstige bezwaren. Wordt consensus als epistemisch criterium van een op andere wijze gedefinieerde waarheid gebruikt dan speelt het probleem dat, om de mogelijke validiteit van het criterium te behouden, het niet daadwerkelijk gebruikt mag worden. Consensus als criterium is in bepaalde zin zelf-refererend. Consensus kan informatief zijn indien het overeenstemming tussen onafhankelijk opererende individuen betreft. Als zodanig is de instemming van iemand met een heersende consensus redundant. Toepassing van het consensus-criterium, wanneer dit leidt tot de aanpassing van iemands opinie, leidt tot dit soort redundantie. De consensus bestaat slechts gedeeltelijk uit de overeenstemming tussen onafhankelijke individuen. Het surplus van de overeenstemming is dan geen indicatie voor de validiteit van de consensus-positie. Het consensus-criterium is zo inconsequentieel – niemand laat zich bij het vormen van

een opinie leiden door het criterium – of invalide. Consensus heeft geen rol in een normatieve methodologie. Niet als doel, niet als middel, en niet als criterium.