Statistical Thinking in Empirical Enquiry
Author(s): C. J. Wild and M. Pfannkuch
Source: *International Statistical Review / Revue Internationale de Statistique*, Vol. 67, No. 3 (Dec., 1999), pp. 223-248
Published by: International Statistical Institute (ISI)
Stable URL: http://www.jstor.org/stable/1403699
Accessed: 03/01/2011 17:46

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=isi.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

International Statistical Institute (ISI) is collaborating with JSTOR to digitize, preserve and extend access to *International Statistical Review / Revue Internationale de Statistique.*

http://www.jstor.org

International Statistical Review (1999), **67**, 3, 223–265, Printed in Mexico
© International Statistical Institute

Statistical Thinking in Empirical Enquiry

C.J. Wild and M. Pfannkuch

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

Summary

This paper discusses the thought processes involved in statistical problem solving in the broad sense from problem formulation to conclusions. It draws on the literature and in-depth interviews with statistics students and practising statisticians aimed at uncovering their statistical reasoning processes. From these interviews, a four-dimensional framework has been identified for statistical thinking in empirical enquiry. It includes an investigative cycle, an interrogative cycle, types of thinking and dispositions. We have begun to characterise these processes through models that can be used as a basis for thinking tools or frameworks for the enhancement of problem-solving. Tools of this form would complement the mathematical models used in analysis and address areas of the process of statistical investigation that the mathematical models do not, particularly areas requiring the synthesis of problem-contextual and statistical understanding. The central element of published definitions of statistical thinking is "variation". We further discuss the role of variation in the statistical conception of real-world problems, including the search for causes.

Key words: Causation; Empirical investigation; Statistical thinking framework; Statisticians' experiences; Students' experiences; Thinking tools; Variation.

1 Introduction

> *"We all depend on models to interpret our everyday experiences. We interpret what we see in terms of mental models constructed on past experience and education. They are constructs that we use to understand the pattern of our experiences." David Bartholomew (1995).*

"All models are wrong, but some are useful" George Box

This paper abounds with models. We hope that some are useful!

This paper had its genesis in a clash of cultures. Chris Wild is a statistician. Like many other statisticians, he has made impassioned pleas for a wider view of statistics in which students learn "to think statistically" (Wild, 1994). Maxine Pfannkuch is a mathematics educator whose primary research interests are now in statistics education. Conception occurred when Maxine asked "What *is* statistical thinking?" It is not a question a statistician would ask. Statistical thinking is the touchstone at the core of the statistician's art. But, after a few vague generalities, Chris was reduced to stuttering.

The desire to imbue students with "statistical thinking" has led to the recent upsurge of interest in incorporating real investigations into statistics education. However, rather than being a precisely understood idea or set of ideas, the term "statistical thinking" is more like a mantra that evokes things understood at a vague, intuitive level, but largely unexamined. Statistical thinking is the statistical incarnation of "common sense". "We know it when we see it", or perhaps more truthfully, its *absence* is often glaringly obvious. And, for most of us, it has been much more a product of experience, war stories and intuition than it is of any formal instruction that we have been through.

There is a paucity of literature on statistical thinking. Moore (1997) presented the following list of the elements of statistical thinking, as approved by the Board of the American Statistical Association (ASA) in response to recommendations from the Joint Curriculum Committee of the ASA and the Mathematical Association of America: *the need for data; the importance of data production; the omnipresence of variability; the measuring and modelling of variability*. However, this is only a subset of what the statisticians we have talked to understand by "statistical thinking" or "thinking statistically". In the quality (or more properly, process and organisational improvement) area, much has been written, but addressing a specific audience. Snee (1990, p. 118) defined statistical thinking as "*thought processes, which recognise that variation is all around us and present in everything we do, all work is a series of interconnected processes, and identifying, characterising, quantifying, controlling, and reducing variation provide opportunities for improvement*". (See also Britz *et al.*, 1997; Mallows, 1998; and Dransfield et al. 1999).

The usual panacea for "teaching" students to think statistically is, with apologies to Marie-Antoinette, "let them do projects". Although this enables students to experience more of the breadth of statistical activity, experience is not enough. The cornerstone of teaching in any area is the development of a theoretical structure with which to make sense of experience, to learn from it and transfer insights to others. An extensive framework of statistical models has been developed to deal with technical aspects of the design and analysis that are applicable once the problem and variables have been defined and the basic study design has been decided. An enormous amount of statistical thinking must be done, however, before we ever reach this stage and in mapping between information in data and context knowledge throughout the whole statistical process. We have little in the way of scaffolding to support such thinking (see Mallows, 1998). Experience in the quality arena and research in education have shown that the thinking and problem solving performance of most people can be improved by suitable structured frameworks (Pea, 1987, p. 91; Resnick, 1989, p. 57).

The authors have begun trying to identify important elements from the rich complexity of statistical thinking. In addition to the literature and our own experience, our discussion draws upon intensive interviews with students of statistics and practising professional statisticians. One set of eleven students, referred to as "students" were individually given a variety of statistically based tasks ranging from textbook-type tasks to critiquing newspaper articles in two one hour sessions. They were interviewed while they solved the problems or reacted to the information. Another set of five students, referred to as "project students" were leaders of groups of students doing real projects in organisations which involved taking a vaguely indicated problem through the statistical enquiry cycle (see Fig. 1(a)) to a solution that could be used by the client. Each was interviewed for one hour about their project. The six professional statisticians were interviewed for ninety minutes about "statistical thinking" and projects they had been involved in. The "project students" and statisticians interviews were structured around the statistical enquiry cycle and were in the form of a conversation which reflected on their approach and thinking during the process of an investigation. This paper is not a report on this particular research (that is being done elsewhere, e.g. Pfannkuch, 1996, 1997), but an attempt to synthesise a more comprehensive picture from these interviews and the literature.

We are not concerned with finding some neat encapsulation of "statistical thinking". Our concerns are deeper than this. We are investigating the complex thought processes involved in solving real-world problems using statistics with a view to improving such problem solving. We are thus interested in developing a framework for thinking patterns involved in problem solving, strategies for problem solving, and the integration of statistical elements within the problem solving. We do not address the thinking involved in developing new statistical methodology and theory. We recognise that much statistical thinking can beneficially take place in day-to-day activities, particularly in the interpretation of information in media and other reports. In interpreting reports, we recognise the applicability of parts of our statistical knowledge about the production, behaviour and analysis of data to the type of information we are receiving and are thus able to critically appraise aspects of that information. The

type of thinking required is very similar, if not identical, to fragments of the thinking performed by someone involved in an enquiry. We see the enquiry cycle as providing a coherent structure that links the fragments and, thus, as an ideal place to start. In subsequent work, we have been specialising and further developing the ideas given here for interpreting statisticial information in reports.

This discussion is organised into a statistical thinking framework for empirical enquiry in Section 2. Section 3 explores "variation". It looks at statistical approaches to real-world problems from the starting point of omnipresent variation. Section 4 takes lessons learned in Section 2 and gives a fragment of a thinking tool for improving investigative skills. Section 5 contains a discussion.

2 A Framework for Statistical Thinking in Empirical Enquiry

Applied statistics is part of the information gathering and learning process which, in an ideal world, is undertaken to inform decisions and actions. With industry, medicine and many other sectors of society increasingly relying on data for decision making, statistics should be an integral part of the emerging information era. Statistical investigation is used to expand the body of "context" knowledge. Thus, the ultimate goal of statistical investigation is *learning* in the *context* sphere. Learning is much more than collecting information, it involves synthesising the new ideas and information with existing ideas and information into an improved understanding.

From the interviews we have built up the four dimensional framework shown in Fig. 1 which seeks to organise some of the elements of statistical thinking during data-based enquiry. The thinker operates in all four dimensions at once. For example the thinker could be categorised as currently being in the planning stage of the Investigative Cycle (Dimension 1), dealing with some aspect of variation in Dimension 2 (Types of Thinking) by criticising a tentative plan in Dimension 3 (Interrogative Cycle) driven by scepticism in Dimension 4 (Dispositions). Who is doing this thinking? Anyone involved in enquiry, either individually or as a member of a team. It is not peculiar to statisticians, although the quality of the thinking can be improved by gaining more statistical knowledge.

2.1 Dimension One: The Investigative Cycle

The first dimension in Fig. 1(a) concerns the way one acts and what one thinks about during the course of a statistical investigation. We have adapted the PPDAC model (Problem, Plan, Data, Analysis, Conclusions) of MacKay & Oldford (1994). The elements of this model should be self-explanatory to statisticians. The statisticians we interviewed were particularly interested in giving prominence to the early stages of PPDAC, namely, to grasping the dynamics of a system, problem formulation, and planning and measurement issues (see Pfannkuch & Wild, 1998).

A PPDAC cycle is concerned with abstracting and solving a statistical problem grounded in a larger "real" problem. Most problems are embedded in a desire to change a "system" to improve something. Even ostensibly curiosity-driven research is usually justified by the idea that the accrued understanding will have long term practical benefits. A knowledge-based solution to the real problem requires better understanding of how a system works and perhaps also how it will react to changes to input streams, settings or environment. Certain learning goals must be met to arrive at the desired level of understanding. A PPDAC investigative cycle is set off to achieve each learning goal. Knowledge gained and needs identified within these cycles may initiate further investigative cycles. The conclusions from the investigations feed into an expanded context-knowledge base which can then inform any actions.

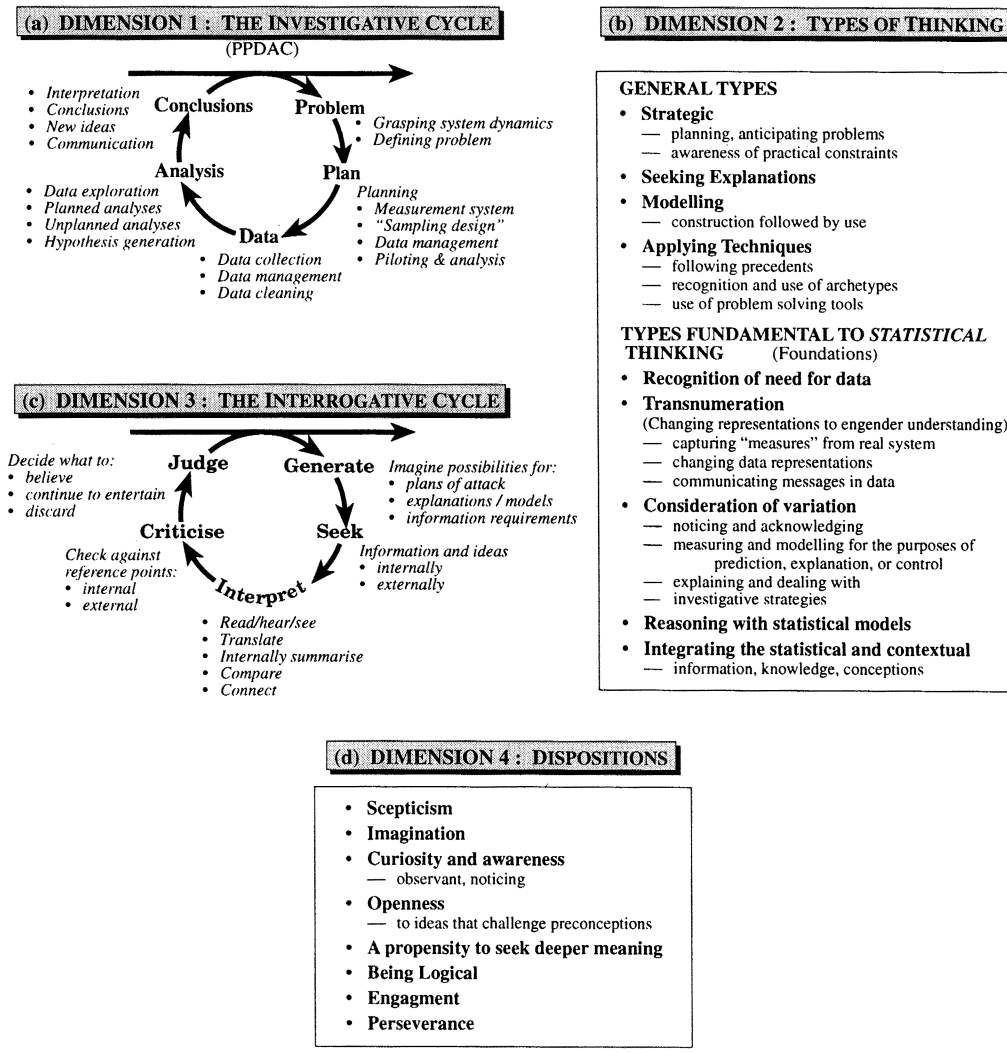

Figure 1. *A 4-dimensional framework for statistical thinking in empirical enquiry*

2.2 Dimension Two: Types of Thinking

A number of types of thinking emerged from the statisticians' interviews and were subsequently refined and modified when we applied them to the student and project-student interviews. The resulting categories are shown in Fig. 1(b). Some types of thinking were common to all problem solving. We will relate these general types of thinking to the statistical context in Section 2.2.2. First, however, we concentrate on types of thinking that are inherently statistical.

2.2.1 Types fundamental to statistical thinking

The types of thinking categorised under this heading in Fig. 1(b) are, we believe the foundations on which *statistical* thinking rests.

Recognition of the need for data: The recognition of the inadequacies of personal experiences and anecdotal evidence leading to a desire to base decisions on deliberately collected data is a statistical impulse.

Transnumeration: The most fundamental idea in a statistical approach to learning is that of forming and changing data representations of aspects of a system to arrive at a better understanding of that system. We have coined the word *transnumeration* to refer to this idea. We define it as "numeracy transformations made to facilitate understanding". Transnumeration occurs when we find ways of obtaining data (through measurement or classification) that capture meaningful elements of the real system. It pervades all statistical data analysis, occurring every time we change our way of looking at the data in the hope that this will convey new meaning to us. We may look through many graphical representations to find several really informative ones. We may re-express the data via transformations and reclassifications looking for new insights. We might try a variety of statistical models. And at the end of the process, transnumeration happens yet again when we discover data representations that help convey our new understandings about the real system to others. Transnumeration is a *dynamic* process of changing representations to engender understanding. Mallows (1998, Section 2) would appear to be advancing a similar idea.

Variation: Thinking which is statistical, in the modern sense anyway, is concerned with learning and decision making under uncertainty. Much of that uncertainty stems from omnipresent variation. The ASA resolution, and Moore and Snee's discussions of statistical thinking all emphasise the importance of variation. The last element of the list following "variation", namely "*for the purposes of explanation, prediction, or control*" is in the original statement of Snee (1990), albeit with a process-improvement spin, but has been dropped from the ASA statement. It is a critical omission. We do not measure and model variation in a vacuum. The purpose influences the way in which it is done. Our concerns with variation also extend beyond "measuring and modelling" to investigative strategies such as randomisation and blocking. In Section 3, we consider the variation theme in much greater detail.

A distinctive set of models: All thinking uses models. The main contribution of the discipline of statistics to thinking has been its own distinctive set of models, or frameworks, for thinking about certain aspects of investigation in a generic way. In particular, methods for study design and analysis have been developed that flow from mathematical models which include random components (see Mallows, 1998). Recently, however, there is a growing desire (enlisting a phrase from David Moore) to nudge "statistics a little further back towards its roots in scientific inference". Large parts of the investigative process, such as problem analysis and measurement, have been largely abandoned by statisticians and statistics educators to the realm of the particular, perhaps to be developed separately

within other disciplines. However, there are more valuable generic lessons that can be uncovered about these parts of the investigative process using other modelling tools. There is a need to expand the *reach* of our statistical models.

Context knowledge, statistical knowledge and synthesis: The raw materials on which statistical thinking works are statistical knowledge, context knowledge and the information in data. The thinking itself is the synthesis of these elements to produce implications, insights and conjectures. One cannot indulge in statistical thinking without some context knowledge. The arid, context-free landscape on which so many examples used in statistics teaching are built ensures that large numbers of students never even see, let alone engage in, statistical thinking. One has to bring to bear all relevant knowledge, regardless of source, on the task in hand, and then to make connections between existing context-knowledge and the results of analyses to arrive at meaning. Ideally, all of this knowledge would be resident in the same brain, but this is often not possible. Major investigations are team efforts which bring together people of differing expertise. Fig. 2 emphasises the synthesis of ideas and information from the context area and from statistics.

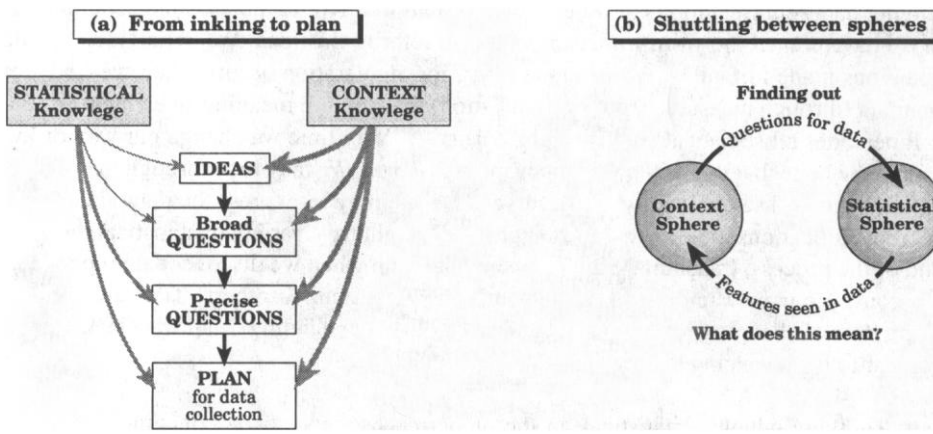

Figure 2. *Interplay between context and statistics*

Fig. 2(a) traces the (usual) evolution of an idea from earliest inkling through to the formulation of a statistical question precise enough to be answered by the collection of data, and then on to a plan of action. The earliest stages are driven almost entirely by context knowledge. Statistical knowledge contributes more as the thinking crystallises. Fig. 2(b) illustrates the continual shuttling backwards and forwards between thinking in the context sphere and the statistical sphere. This goes on all the time throughout PPDAC. For example, at the analysis stage questions are suggested by context knowledge that require consulting the data—which temporarily pushes us into the statistical sphere—whereupon features seen in the data propel us back to the context sphere to answer the questions, "Why is this happening?" and "What does this mean?"

2.2.2 General types of thinking applied in a statistical context

Strategic thinking

By strategic thinking, we mean thinking aimed at deciding upon what we will do (next or further into the future) and how we will do it. This includes such things as: planning how to attack a task; breaking tasks down into subtasks; setting deadlines for subtasks; division of labour; and anticipating problems and planning to avoid them. An important part of strategic thinking is having an awareness of the constraints one is working under and taking them into account in planning.

Real statistics is less about the pursuit of the "correct" answer in some idealistic sense than about doing the best one can within constraints. Many factors limit the quality and effectiveness of the thinking. Some of these factors are internal to the thinker. *Lack of knowledge* obviously constrains thinking. Unfortunately, what we "know" is not only our greatest asset but also our biggest curse because the foundations of what we "know" are often not soundly based. Our *preconceptions* can lead us astray in many ways, for example, by blinding us to possibilities because what we "know" determines where we look, and by desensitising us to important information. The challenging of something we "know" and take for granted can remove an obstacle and lead to new insight. This often occurs when people with different backgrounds discuss the same problem. Consulting statisticians see it at work in their clients when a quite innocent question surprises the client, loosens a previously held preconception, and leads to the client seeing the problem in a new way. We tend to solve problems by following "precedents". In applied research, this happens all the time and often the statistical methods of the *precedents are inadequate*. As far as *dispositions* (Dimension 3) are concerned, someone who is not curious, imaginative, sceptical and engaged will be less effective than someone who is. There is also an *ability* factor operating. Faced with the same set of information, some people will be better at making useful connections and grasping the essential features than others. And *inadequate communication skills* limit the ability to extract vital information and ideas from clients and others.

Other constraints are due to the environment the thinker is operating in. These include the general *time, money and materials* constraints, the imperfection of all human communication which results in misunderstandings and gaps in transmission of essential knowledge, and *limitations of the data* available. Very often, the problem we would like to solve is simply not soluble on the basis of the information we can get. For example, it may be impossible to capture with feasible measurement processes the characteristics we would like to capture. It may be impossible to sample the desired population or even a good approximation to that population, and so on.

This paragraph relates to particular constraints faced by statistical consultants, but students and other researchers are subject to some closely related constraints. The consultant works on problems owned by someone else. In other words, the statistician is in the position of having to satisfy "clients". This brings additional constraints which run deeper than time-and-materials constraints. Major decisions are made by, or must be cleared with, the client. The problem territory tends to be mapped out and even ring-fenced by the client. The client is often the chief source of context information so the statistician is not only constrained by the quality of communication and the extent of the client's knowledge, but will also tend to take on board the client's preconceptions. As the client is the final arbiter, the statistician is constrained by what the client can understand and accept. This can be strongly influenced by a number of what might be described as *psychological factors*. Statisticians have to gradually build up the client's trust in their judgement and abilities. An important consideration in "building trust" is not taking clients too far from territory in which they feel secure. An important element in client security is, in the words of colleague Chris Triggs, "what has been done in the field before". We call this the *first-in-the-field effect*. Early work in a field tends to take on an authority of its own whether or not it is warranted. It can influence every decision in the investigative process, right through to presentation. A related psychology of measurement effect

concerns the *sanctity of the measured variable*. To many clients, the way in which a variable has been measured takes on a meaningfulness and inviolability that a statistician might disregard, given the arbitrary elements in the initial choice of the variable. (This is not universal. Some client groups such as engineers are very sophisticated in this area.) The use of transformations in analysis is an area in which these issues come into sharp focus. Pfannkuch & Wild (1998) give a much more detailed and wide-ranging discussion, derived from the statisticians' interviews, of the realities of working with clients.

Modelling

Constructing models and using them to understand and predict the behaviour of aspects of the world that concern us seems to be a completely general way of thinking. All models are oversimplifications of reality in which information is necessarily discarded. We hope that we have caught the essential features of a situation and the loss of information does not invalidate our conclusions. Fig. 3 illustrates the way in which we learn about the context reality as a statistical investigation proceeds. As our initial quotation from David Bartholomew makes clear, "understanding" builds up in mental models of the context reality. These models are informed by information from the context reality, e.g. incorporating "expert knowledge". In an ideal world, we would be continually checking the adequacy of the mapping between model and reality by "interrogating" the context reality. Some of the information we seek and get from the context reality is statistical data. We build statistical models to gain insights from this information ("interpret") which feed back into the mental model. "Statistical models" here is more general than something like logistic regression. It refers to all of our statistical conceptions of the problem that influence how we collect data about the system and analyse it. Fig. 3 also incorporates the role of statistical knowledge and experience. Most obviously, it is a major determinant of the statistical conceptions we form in order to obtain and analyse data. Additionally, depending on the problem and the education and experience of the thinker, statistical elements can also be part of the way we think about the world and thus be integral parts of our mental models of the context reality.

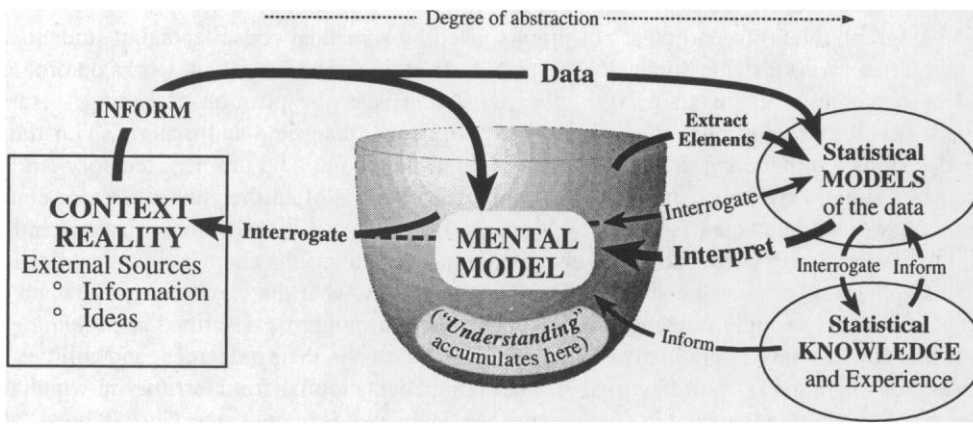

Figure 3. *Learning via statistics*

Applying techniques

A basic problem solving technique in the mathematical sciences is to find a way of mapping a new problem onto a problem that has already been solved so that the previously devised solution can be applied or adapted. The whole discipline of statistics is itself a manifestation of this strategy. Statistical theory makes the mapping process efficient by creating problem archetypes and linking them to methods of solution. To use statistics, we first recognise elements of our context that can be usefully mapped onto a model (a process of abstraction from the particular to the generic), operate within that model, and then we map the results back to context (from the generic to the particular). (Additionally, applied statisticians are always borrowing problem-solving ideas from previous experience with other problems and other data sets.)

Implementation of the problem-archetype strategy, and indeed the practical application of any technique, algorithm or concept, involves the three steps shown in Fig. 4. Instruction tends to focus on step 2, mechanical application. However, steps 1 (recognition) and 3 (interpretation in context) are: first, vital to step 2 having any utility, and second, inordinately more difficult. This is particularly true for the recognition step. (The project students needed to make constant external checks with their supervisor about whether they were on the right track.) One can deal with the mechanics of procedures by simply talking about them, establishing them with a few exercises and then moving on. Synthesis, insight, critical thinking and interpretation happen in the realm of the particular and require exposure to large numbers of disparate situations (cf. Wild, 1994).

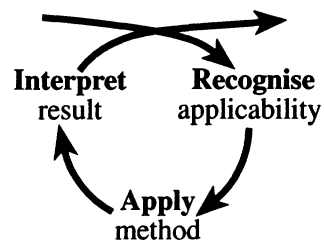

Figure 4. *Using any technique*

It is characteristic of statistics that we apply relatively sophisticated statistical models to the analysis of data and experimental design. Of all our statisticians, however, only the one operating in quality improvement seemed to use tools (e.g. cause-and-effect diagrams and process-mapping techniques) to analyse the nature of the problem itself. For the others, it seemed to be a process of imaginative construction of a mental model of the system, without discernible organisation. (The type of thinking "***seeking explanations***" has not been discussed in this section, but will be an important theme in Section 3.)

2.3 Dimension Three: The Interrogative Cycle

The Interrogative Cycle illustrated in Fig. 1(c) is a generic thinking process in constant use in statistical problem solving. From a detailed analysis of the project-students' and students' transcripts, it appears that the thinker is always in one of the interrogative states while problem solving. The cycle applies at macro levels, but also at very detailed levels of thinking because the interrogative cycle is recursive. Subcycles are initiated within major cycles, e.g. the "checking" step of any cycle can initiate a full interrogative subcycle. The ordered depiction on a wheel is an idealisation of

what perhaps should happen. In reality steps are often missed. We discuss the Interrogative Cycle as we observed it, being applied to statistical enquiry and statistical critique. The "thinker" is anyone involved in these activities. We now explore the components in more detail.

Generate: By this we mean imagining and brainstorming to generate possibilities, as an individual or in a group. We might be applying this to a search for possible causes, explanations and mechanisms, to the ways parts of a system might interrelate and to other building blocks of mental and statistical models. We might be applying it to the types of information we need to seek out to fill an information gap or to check out an idea, or to plan an approach to a problem or subproblem. The generation of possibilities may be from the context, the data or statistical knowledge and apply to the present problem, or may be registered for future investigation (hypothesis generation).

Seek: Generation tends to be followed by a seeking or recalling of information. This may be internal or external. For internal seeking, we observe people thinking "I know something about this" and digging in their memories for the relevant knowledge. External seeking consists of obtaining information and ideas from sources outside the individual or team. Working statisticians talk to other people about their problems—clients, colleagues, context-matter experts, people "working in the system". Seeking includes reading relevant literature. At the macro level it includes the collecting of statistical data, while at a more detailed level it includes querying the data in hand.

Interpret: By this we mean taking and processing the results of our seeking.

Read/see/hear ⟶ Translate ⟶ Internally summarise ⟶ Compare ⟶ Connect

This process applies to all forms of information including graphs, summaries and other products of statistical analysis. "Connect", the endpoint of "interpret" refers the interconnecting of the new ideas and information with our existing mental models and enlarging our mental models to encompass these interrelationships. Some of the problems observed in student thinking involved making one connection and then rushing to "judge" rather than trying to make multiple connections or going through the criticism phase.

Criticise: The criticism phase applied to incoming information and ideas involves checking for internal consistency and against reference points. We ask, "Is this right?" "Does this make sense?" "Does this accord with what else I or others know?" We check against *internal* reference points—arguing with ourselves, weighing up against our context knowledge, against our statistical knowledge, against the constraints we are working under, and we anticipate problems that are consequences of particular choices. We may also check against *external* reference points such as: other people (i.e. talk to clients, colleagues, experts, "workers in the system"); available literature and other data sources (e.g. historical data).

We can similarly try to take a mental step back and monitor our own thinking. Educational theorists talk about metacognition, of recognising and regulating one's normal modes of thought (see Shaughnessy, 1992). Reference points to check against here include the following: (1) *The purpose of the thinking*: for example, "Does this address the question the client wants answered?", or some sort of agreed objectives. (2) *Belief systems*: "Am I being unduly guided by unwarranted preconceptions—my own, my client's, or my community's?" Pfannkuch & Wild (1998) have some good cautionary tales from the experiences of our statisticians. (3) *Emotional responses*: One of our project students was worried about how the company's treatment of her seemed to be influencing the way she was approaching the problem and viewing the data.

Judge: This is the decision endpoint of criticism. What we keep, what we discard or ignore, what

we continue to tentatively entertain, what we now believe. We apply *judgement* to such things as: the reliability of information; the usefulness of ideas; the practicality of plans; the "rightness" of encapsulation; conformance with both context-matter and statistical understanding; the relative plausibility of competing explanations; the most likely of a set of possible scenarios; the need for more research; and the many other decisions involved in building and reasoning from models.

The result of engaging in the interrogative process is a ***distilling and encapsulating*** of both ideas and information. Internal interrogative cycles help us extract essence from inputs, discarding distractions and detail along the way (Fig. 5).

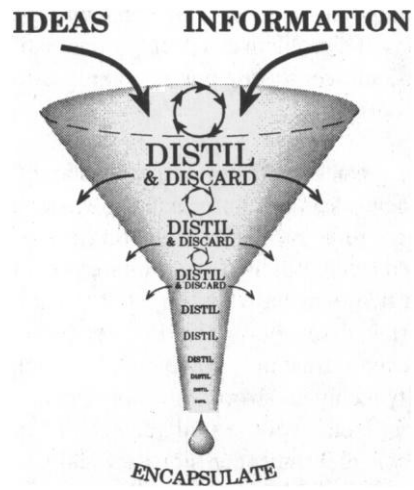

Figure 5. *Distillation and encapsulation*

2.4 Dimension Four: Dispositions

In this subsection, we discuss personal qualities categorised in Fig. 1(d) which affect, or even initiate, entry into a thinking mode. The nature of these dispositions emerged from the statisticians' interviews and we could subsequently recognise them at work in the students. We think these elements are generic, but again we discuss them as we observed them—in the context of statistical problem solving.

Curiosity and Awareness: Discoveries are triggered by someone noticing something and reacting to internal questions like "Why?", or "How did that happen?", or "Is this something that happens more generally?", or "How can I exploit this?" Being observant (aware) and curious are the well-springs of the question generation process that all innovative learning results from. Wild (1994) formed the slogan "Questions are more important than answers" to emphasise this point. Statistician Peter Mullins stressed the importance of "*noticing variation and wondering why*" for generating ideas for improving processes and service provision. We hazard that this very basic element of statistical thinking is actually at the root of most scientific research. "Noticing and asking why" is also critical for successful data exploration and analysis.

This brings us to ***engagement***. When the authors become intensely interested in a problem or area, a heightened sensitivity and awareness develops towards information on the peripheries of our experience that might be related to the problem. We suggest that this experience is fairly general.

People are most observant in those areas that they find most interesting. Engagement intensifies each of the "dispositional" elements curiosity, awareness, imagination and perseverance. How do we become engaged? Spontaneous interest is innate. Background knowledge helps—it is hard to be interested in something one knows nothing about. Being paid to do a job helps, as does the problem being important to people we care about. This may be our main difficulty in getting statistics students to think. They simply do not find the problems they are asked to think about interesting enough to be really engaged by them. We observed the effects on performance of engagement with some tasks and not others in the statistics students.

Imagination: It is hard to overemphasise the importance of imagination to statistical thinking. This is somewhat ironic given popular stereotypes of statisticians. The formation of mental models that grasp the essential dynamics of a problem is a deeply imaginative process, as is viewing a situation from different perspectives, and generating possible explanations or confounding explanations for phenomena and features of data.

Scepticism: By scepticism, we mean a tendency to be constantly on the lookout for logical and factual flaws when receiving new ideas and information. It is a quality all our statisticians both possess and value. Some writers refer to this as "adopting a critical attitude". Gal *et al.* (1995) and Pfannkuch (1996) discussed critical thinking in the interpretation of statistically based reports and media articles. Scepticism here was basically targeted towards, "Are the conclusions reached justified?" There may be worries about the motivation, predispositions and objectiveness of the writer which would effect the level of trust in anything that had been done. Experienced statisticians are likely to evoke automatically technical "worry questions" concerning the appropriateness of the measurements taken, the appropriateness of the study design, the quality of the data, the suitability of the method of analysis, and whether the conclusions are really supported by the data. Postulated explanations create worries about whether this really is the only plausible explanation.

Another aspect involves *a sense of number* and scepticism. A precursor step towards "Is this information/conclusion *justified*?" is "Is this information/conclusion even *credible*?" One of our statisticians told the simple story of reported attendance rates at a free outdoor concert in Auckland. If the figures were correct, that would mean that one in every three Aucklanders, one in nine New Zealanders, would have needed to have attended and that was, frankly, incredible. The information is discounted at this first hurdle. However it should be noted that one is much less inclined to be sceptical when conclusions fit one's own preconceptions. A conscious effort may be required to counter this.

Being logical: The ability to detect when one idea follows from another and when it does not, and to construct a logical argument is clearly important to all thinking. Synthesis of new information with existing knowledge is largely a matter of seeing *implications*. Logical reasoning is the only sure way to arrive at valid conclusions. To be useful, scepticism must be supported by an ability to reason from assumptions or information to implications that can be checked against data.

A propensity to seek deeper meaning means not simply taking things at face value and being prepared to dig a little deeper. Of the other "dispositions", ***openness*** helps us to register and consider new ideas and information that conflict with our own assumptions and ***perseverance*** is self evident.

Can "dispositions" be taught?

Schoenfeld (1983) analysed the mathematical problem solving experience within individuals in terms of a "manager" and an "implementer" working in tandem. The manager continually asks questions of a strategic and tactical nature deciding at branch points such things as which perspective to adopt and which direction to take or abandon. We have described the characteristics above as "dispositions". They tend to initiate manager functions. We first thought of the dispositions as innate

characteristics of the thinker but had to modify this with the idea of "engagement". A person's "dispositions" are problem dependent—they change according to the degree to which the person is engaged by the problem. One of our statisticians was adamant that some people are sceptical, others are credulous, and there is little one can do about it. The authors are less pessimistic. It seems to us that credulousness in a particular area is a result of ignorance. As you gain experience and see ways in which certain types of information can be unsoundly based and turn out to be false, you become more sceptical. Moreover, all we want in operational terms from scepticism is a prompting to raise certain types of "worry" question [cf. Gal *et al.*'s (1995)] concerning the reliability of information, which can be taught (see Section 4).

3 Variation, Randomness and Statistical Models

3.1 Variation as the Starting Point

The centrepiece of the quality and ASA definitions of statistical thinking is "variation" or "variability". Any serious discussion of statistical thinking must examine the role of "variation". The "variation" terminology and message seem to have arisen in one small area of statistical application, namely that of quality, and their penetration into other areas would appear to be slight. If "variation" (as a major source of uncerainty) is indeed to be the standard about which the statistical troops are to rally, we need to arrive at a common conception of statistics in terms of "variation". This section attempts such a conception. Moreover, we are striving for a view of statistics "from the outside".

The first three "variation" messages are: variation is omnipresent; variation can have serious practical consequences; and statistics give us a means of understanding a variation-beset world. Subsequent messages concern how statistics goes about doing that.

Omnipresence: Variation is an observable reality. It is present everywhere and in everything. Variability affects all aspects of life and everything we observe. No two manufactured items are identical, no two organisms are identical or react in identical ways. In fact, individual organisms are actually systems in constant flux. The aforementioned refers only to real variation inherent in the system. Fig. 6 depicts how, when we collect data from a system, this real variation is supplemented by variation added in various ways by the data collection process.

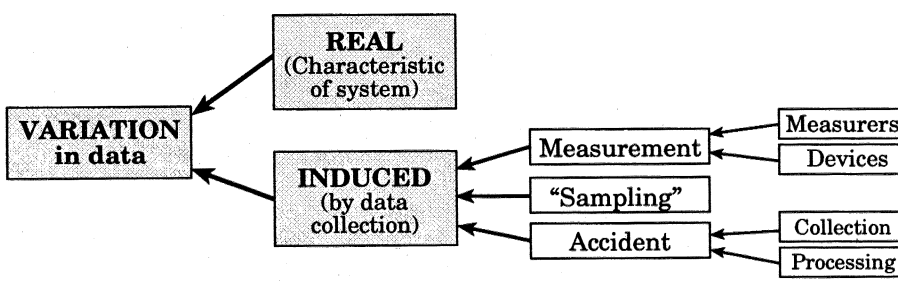

Figure 6. *Sources of variation in data*

Practical impact: Having established that variation is everywhere, we have then to demonstrate the important practical impacts of this variation on peoples' lives and the way they do business. It is variation that makes the results of actions unpredictable, that makes questions of cause and effect difficult to resolve, that makes it hard to uncover mechanisms. Variation is the reason why people

have had to develop sophisticated statistical methods to filter out any messages in data from the surrounding noise.

3.2 *Predict, Explain and Control*

Fig. 7 categorises rational responses to variation in a system in the world of action. This is idealistic. The way people actually *do* react to variation can be quite another story! (See Joiner, 1994, Chapter 10).

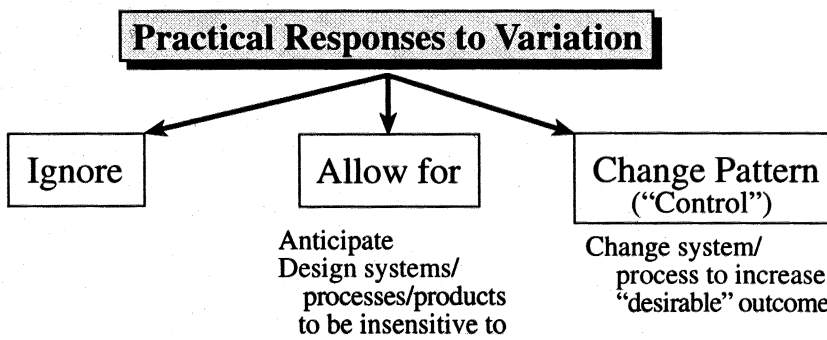

Figure 7.

First, we can pretend that the variation does not exist, e.g. behave as though every object or organism is the same or differs in some deterministically known way. In some circumstances this works admirably. If it did not we would have to write off all of applied mathematics and every field it fertilises. Second, we can investigate the existing pattern of variation and come up with ways of working around it as in our system of clothing and shoe sizes. Variation is allowed for at the design stage in quality management approaches to manufacturing where one wishes to design a product that is "rugged" or "robust" to the variability of uses to which it will be put and conditions to which it will be subjected. Third, we can try to change the pattern of variation to something more desirable, e.g. to increase average crop yield or reduce a death rate. We do this by isolating manipulable causes, or by applying external treatments. The former approach is often used in quality improvement or in public health, the latter is frequently used in agriculture or in medical research aimed at the treatment of individual patients.

Statisticians model variation for the purposes of prediction, explanation, or control. *Control* is changing the pattern of variation to something more desirable. *Prediction* is the crucial informational input to "Allow for" in Fig. 7. *Explanation*, gaining some level of understanding of why different units respond differently, improves our ability to make good predictions and it is necessary for control. Causal and mechanistic explanation is the goal of basic (as opposed to applied) science. As soon as we ask "Why?", we are looking for causes. While on the one hand variation may obscure, it is the uncontrolled variation in a system that typically enables us to uncover causes. We do this by looking for patterns in the variation. Fig. 8 picks up this idea in a way that relates back to the goals in Fig. 7.

Statisticians look for sources of variability by looking for patterns and relationships between variables ("regularities"). If none are found, the best one can do is estimate the extent of variability

and work around it. Regularities may or may not correspond to causes. In terms of solving practical problems, causes that cannot be manipulated are operationally equivalent to any other observed regularity, although they will give us more confidence in our predictions. The presence of regularities enables us to come up with predictions and measures of variability that are more locally relevant, e.g. more relevant to an individual patient. Manipulable causes open the option of control.

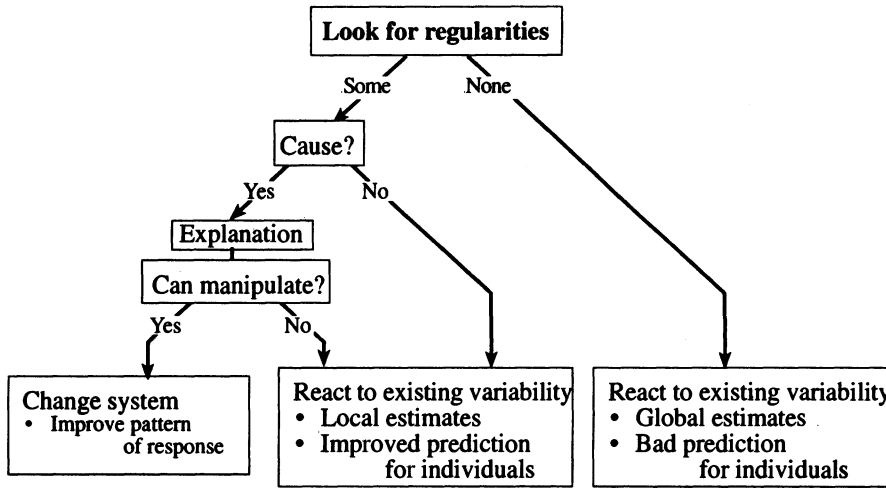

Figure 8.

3.3 The Quest for Causes

In our research, two groups of first-year students were given media clippings and similar items containing statistical information, and then interviewed individually about their responses. Initially, our approach to the student transcripts was that of teachers marking term papers, of looking for mistakes and gaps, for what the students had "done wrong". One item was based on Tversky & Gilovich (1989). The streaks that sports fans see in sports data, and then proffer all sorts of causal explanations for (e.g. falters under pressure), can often be explained entirely in terms of a random, e.g. binomial, model (see also Moore, 1990; Falk & Konold, 1992; Biehler, 1994 presents other perspectives). The item run concerned a basketball player with a 70% success rate of free throws succeeding with only 2 out of 5 throws. Under a binomial model, this is not a particularly unusual event. We think of this as "the statistics teachers' point". Our students proffered all sorts of causal explanations. As statistics teachers, we thought they had missed the point. Mark that one wrong! For the next group, we loaded the item entirely in favour of the statistics teacher's point: "*The team manager attributed her performance to normal variation, that she scored 70% in the long run and that 70% was only an average so that you had to expect some low scores now and again.*" Even then we saw the tip of the deterministic-causal-thinking iceberg. One student said, "*the manager's comments are OK if that is the way he wants to look at the score and not on 'we want to win'* " and then he gave possible causes.

This comment eventually overturned our attitude. The student was right. There is a real problem underlying this item. Coaches and managers of sports teams are seeking to learn from their observations so that they can work on improving player skills and strategy, and better deploy their pool of available players. A random model is of no help at all in this regard. The statistics teacher's concerns

do not replace the need to search for causes and predictors of good and bad performance. It is that search that is of primary importance.

The real problem underlying the statistical problem very often involves searching for and isolating causes of some response. We have run a variety of stories including medical stories and prison suicides. Whenever students have contextual knowledge about a situation, and their experience has given them some notion of the nature of the real problem, they will come up with a range of possible causal explanations with little or no prompting. This appears to be a well developed impulse that has not been explicitly taught. It is a first and direct step towards solving the primary problem. The real purpose of many of the new statistical ideas we teach is simply to moderate the search for causes by preventing a premature jumping to conclusions—"Hey, not so fast . . . ." This role is secondary and subtle. It is probably not surprising then, that even after some statistical instruction, the randomness ideas are much weaker in students than the impulse to postulate causes. Probabilistic thinking is not so much an alternative to deterministic thinking, as some statistics educators (Shaughnessy, 1992) and statisticians (Hoerl *et al.*, 1997) have suggested, but something to be grafted on top of the natural thinking modes that directly address the primary problem. As an interesting aside, if an explanation or cause has already been suggested to students for a particular set of data or if the data has been presented stratified in some particular way, it can take a great deal of prompting for the student to go beyond the explanation given, to think that there may be other explanations and start coming up with ideas. This latter is a quite different incarnation of "Hey, not so fast, ..."

What does statistics education have to say about causation? By far the loudest message is, "correlation is not causation". This is the statistician as Cassandra, the harbinger of doom saying "this way lies disaster". True, we usually go on to make the important point that the randomised experiment is the most convincing way of establishing that a mooted relationship is causal. But, as stressed by Holland (1986), Cox (1992) and few others outside of quality and epidemiology, this greatly undersells the true importance of the search for causes. Solving most practical problems involves finding and calibrating change agents. Statistics education should really be telling students something every scientist knows, "The quest for causes is the most important game in town". It should be saying, "Here is how statistics helps you in that quest. Here are some general strategies and some pitfalls to beware of along the way . . . . . ". It should not just be preventing people from jumping to false conclusions but also be guiding them towards valid, useable conclusions—replacing Cassandra by a favourite literary detective.

Thinking about causes

It is ironic that the uncontrolled variability in a system provides us with the best opportunities to do the detective work that uncovers causes. By checking for relationships between upstream variables and downstream responses, we can identify possible causal factors. Observation precedes experimentation. All *ideas* for possible causal relationships originate in observation, whether anecdotal or from formal studies. And as we continually stress, randomised experiments provide the most convincing way of confirming or refuting the causal nature of an observed relationship.

Conducting any sort of study to detect causes and estimate their effects proceeds from ideas about profitable places to look, ideas which draw almost exclusively on context-matter knowledge and intuition. Ideas about possible causes and other factors that might be important predictors of the behaviour of the response are translated into a set of variables to measure (transnumeration) and data is collected to facilitate investigation of relationships between measured variables and the responses of interest. The primary tools of analysis in the search for causes are models of the regression type, i.e. models for exploring how Y-behaviour changes with changes in X-behaviour. (The humble scatter plot falls into this class.)

Cox (1992) distinguishes between: response variables (those whose behaviour we want to find

causal explanations for); intermediate response variables (which measure intermediate effects that happen along the way from initial state to response state) and explanatory variables (those we want to use to explain or predict the behaviour of the response). Explanatory variables are further categorised into possibly causal variables, intrinsic properties of entities under study, and non-specific (e.g., different countries). Intrinsic variables are those whose values cannot be manipulated. Intrinsic variables are often included to improve our ability to detect relationships, improve the precision of estimation of effects and to explore how a cause may act differently for different types of entity (interactions). Detection of strong relationships between non-specific variables and a response lead to a search for new explanatory variables, variables associated with the non-specific variable which could conceivably explain the response. For example, when disease rates differ greatly between countries, we start looking among factors that differ between the countries as possible causes. We note that the above distinctions between variables are much more than distinctions for analysis. As a set they constitute a general thinking tool which adds clarity to the way in which the context-matter problem is conceived. They are an example of the way in which statistical knowledge or training can feed into a core mental model of the context reality that is understandable by statisticians and non-statisticians alike, the "inform" arrow linking statistical knowledge and mental model in Fig. 3.

Some consequences of complexity

Most real systems are enormously complex with variation in innumerable components, each of which could contribute to the response of interest. We are incapable of handling such complexity and need strategies to "sift" the large numbers of possibilities for a much smaller number of promising leads. The quality control distinction between *special cause* and *common cause* variation can be seen in this light. It gives a means of distinguishing situations (special-cause variation) in which the seemingly instinctive human reaction of looking around for something unusual occurring in the system just prior to the problem event is likely to be a profitable strategy for locating the cause, from situations (common-cause variation) in which this strategy is unlikely to be profitable and may even be harmful.

The main statistical "sifting"-strategy is to restrict attention to variables which have strong associations with the response of interest. We have no hope of identifying a cause and characterising its effects if it acts in complexly different ways for different individuals or at different times. The only causes that we can hope to find are those that act in a reasonably uniform or regular way. Moreover, we will only detect the existence of a cause if we think of some way of looking at the situation that will reveal that regularity (transnumeration). There must be sufficient "natural variability" in a cause-variable in the system for the effect of this variability on the response to be seen. Causal variables that we miss using the above strategy are unlikely to be good agents for making substantial changes unless settings are used that lie far beyond the range of variability seen in that variable in the data.

From association to causation

It is at this point that Cassandra makes her entrance. And the world really does need her warnings. It is clear that people do jump far too quickly to causal conclusions. But "correlation is not causation" is simply a "Hey, not so fast" warning and we need to supply ways of moving on from there. The search process has not given us a set of causes. It has only given us a set of promising contenders for causal status. Our main worry at this point stems from the fact that we have not considered the universe of relevant variables, but just that subset that happened to come to mind. We are worried that other unconsidered factors, those sinister lurking variables of textbook fame, may be producing the relationships we are seeing. Thus, we challenge the causal assumption, whether our own or

somebody else's. We rack our brains for other possible explanations and for strategies for testing these explanations. This behaviour has to be learned. It comes naturally to very few students. The goal of the scientist is to reach a position at which there are no other plausible explanations at the current level of understanding. To do this, we need strategies which use observation, experimentation and analysis to discount all other plausible alternatives.

Where experimentation is not possible and one must make decisions based upon using observational studies, there is a range of ideas about what strengthens the impression that a causal contender is in fact a cause. The criteria of A.B. Hill (1965, see also Gail, 1996) are a good starting point. In epidemiology and quality, the finding of causes with a view to improving systems is not a philosophical problem but a pressing practical imperative. Substantial literatures have grown up in these fields. Cox (1992, 1993) and Holland (1986) also view questions of causation with practical applications clearly in mind. In view of the fundamental importance of the search for causes, there is a real need to synthesise this material into accounts which are more accessible for practising investigators and for teachers.

Levels of "causal proof"

Decisions to take action tend to be made on the basis of a "best guess" in the light of the available information. They seldom wait for incontrovertible evidence of causality. The results can be spectacularly good. Take cot death in New Zealand. Research showed strong relationships between cot-death rates and certain behaviours, e.g. the way the baby was put down to sleep. There was no incontrovertible proof that the behaviours caused cot death but the idea was sufficiently plausible to mount publicity campaigns and the advice given to new mothers by doctors. Cot death rates halved. There is a level of assurance at which decision makers are prepared to take what they consider to be a small chance and take action. There are many factors affecting this level of assurance. The degree of causal proof it takes will probably depend on many factors including the difficulty of making (and reversing) changes to the system, the consequences of making a wrong call, and the number of people who must be convinced before action is taken. We are all brought up on the smoking-cancer debate as the primary example of the difficulties in establishing causality. In that debate, there were (and are) entrenched and powerful vested interests with a high political profile. Not surprisingly, the level of proof required in such circumstances is extremely high. An industrial production manager would have made the call long before, with the greater attendant risk of getting it wrong.

3.4 Modelling Variation

A number of statisticians have told us that the biggest contribution of statistics is the isolation and modelling of "signal" in the presence of "noise". The base problem with statistical data, is how to make some sort of sense of information that is, if one considers the details, of mind-boggling complexity. The main statistical approach to solving this problem begins by trying to find patterns in that data. Context knowledge may give us some ideas about where to look and what to expect. Statistical methodology gives us tools to use in the search. Common experience tells us that studies conducted under very similar conditions always give results which are different in detail, if not in broad thrust—patterns seen in data from one study are never repeated identically in another. The base problem, then, is to come up with strategies for separating phenomena which are "likely" to persist more generally from those that are purely local, to sift the enduring from the ephemeral. Patterns which persist provide the basis for forecasting, control and insight. Statisticians have evolved particular sets of strategies for "solving" this problem—strategies based, in the main, on probabilistic modelling. We often say that an important function of probability models and statistical inference is to counteract a human tendency to "see" patterns where none exist. As statistician (and also zoologist)

Brian McArdle put it so vividly in a personal interview, "*The human being is hard-wired to see a pattern even if it isn't there. It's a survivor trait. It lets us see the tiger in the reeds. And the downside of that is that our children see tigers in the shadows on the wall.*" It is not entirely true that no patterns appear in purely random phenomena. These patterns are real to the brain in the sense that we can recognise features that would help us reproduce them. However, such patterns are (i) ephemeral, and (ii) tell us nothing useful about the problem under study. In other words, they are meaningless. Part of our reasoning from random models is to say that we will not classify any data-behaviour as "enduring" if it closely resembles something that would happen reasonably frequently under a purely random model.

The distinction between "explained" and "unexplained" variation is important here. We generally try to find meaning in explained variation, the patterns which we have not discounted as ephemeral, the "signal". Unexplained variation, or "noise", is what is left over once we have "removed" all the patterns. It is thus, by definition, variation in which we can find no patterns. We *model* unexplained variation as being generated by a (structureless) random process. We have no idea whether this variation really is random; this is not something that bothers us. If random sampling really has occurred, there is an element of randomness in the noise. However, measurement error and components of the variation in the original process typically contribute to the unexplained variation and there is no way of knowing whether these behave randomly or not. In fact, randomness is just a set of ideas, an abstract model, a human invention which we use to model variation in which we can see no pattern. The very physical models we use to illustrate randomness are, with sufficient knowledge, actually deterministic (see Stewart, 1989, Chapter 14). It is all part of an attempt to deal with complexity that is otherwise overwhelming, and it tends to be a model-element of last resort. The level at which we impose randomness in a model is the level at which we give up on the ability to ask certain types of question, questions related to meaning and causation.

Language such as "real or random" or referring to the possibility that "the observed difference is due to chance" actively obscure the distinction between the underlying problem and a statistical approach to its solution. In talking about a project he did on mangroves one student said "*My teacher explained it [t-test] to me that the results I got were due to chance. I still don't think that statement makes any sense. I can understand what chance is when you are rolling a dice. I don't really understand what chance is when you relate it to biological data. Everything you could possibly measure is going to be due to some environmental impact.*"

Some writers in quality have taken to saying, "*all variation is caused*"; e.g. Joiner & Gaudard (1990), Pyzdek (1990). The latter repudiates the "outdated belief that chance causes should be left to chance". These claims seem to be predominantly motivated by concerns about human psychology. Tomorrow, with new information, insight or technology, we may be able to find patterns in what today looks random, to trace causes from those patterns, and to improve the system (Pyzdek gives examples where this has occurred). The propensity to do so may well be lost if the idea is internalised that this variation is "just random". In commenting on the difficulties people have with coming to grips with statistics, Shaughnessy (1992) wrote "*the real world for many people is a world of deterministic causes . . . there is no such thing as variability for them because they do not believe in random events or chance.*" We do not need to ask them to. Variability is a demonstrable reality. Randomness need not relate to any belief system about the true underlying nature of reality. It is simply a response to complexity that otherwise overwhelms us. The unexplained variation may well be the result of "a multiplicity of causes", to use the phrase of Falk & Konold (1992). Few would dispute that much unexplained variability is of this type. But, the statistical response is that if we can see no structure there, we will model it as having been generated randomly.

From these models, we make inferences. We assume that the data has been randomly generated according to the model and use probability as the link between population/process and data. This is the very heart of the statistics we teach. Our models, including their random components, stand or

fall on the practical usefulness of the answers they produce. There are some clear success stories, e.g. the insurance industry. To use models with random components, we have to be able to: first, recognise that such models provide a useful framework for considering the problem; second, build and fit an appropriate model; and third, deduce implications from that model. The third step involves some understanding of how random models behave. There is an emerging literature on the difficulties in gaining that understanding; see, for example, Pfannkuch & Brown (1996), Garfield & Ahlgren (1988), Konold (1994). Our inferential paradigms are also subtle and difficult to grasp, but we will not discuss that here (see Mallows 1998, Section 7; and Cox, 1997, Section 5).

One of the stories we have been showing students, and our reaction to it, niggled at us for a long time. It was a news story about an apparent jump in prison suicides, the sort that leads to accusatory finger pointing and the pushing of different causal explanations by different sectional interests. We automatically did a quick check against Poisson variation. The figure was within reasonable limits. We sensed a tendency, as a consequence of this calculation, not just to disregard the hype, but to disregard the problem. However, prison suicides are an important problem and people should be looking for causes. It took a long time to realise that what the lack of significance really tells us is to adopt the common-cause-variation strategy of in-depth study rather than the (popular) special-cause-variation strategy of looking among recent changes for a cause.

Relating the "variation" words

We conclude this section by putting some summarising organisation into the "variation" terminology. *Special-cause* versus *common-cause* variation is a distinction which is useful when looking for causes, whereas *explained* versus *unexplained* variation is a distinction which is useful when exploring data and building a model for them. An understanding of variation in data could be built on these suppositions: (1) variation is an observable reality; (2) some variation can be explained; (3) other variation cannot be explained on current knowledge; (4) *random* variation is the way in which statisticians model unexplained variation; (5) this unexplained variation may in part or in whole be produced by the process of observation through *random sampling*; (6) randomness is a convenient human construct which is used to deal with variation in which patterns cannot be detected.

4 Thinking Tools

Gal *et al.* (1995) used the term "*worry questions*" when discussing the critical appraisal of reports—questions to invoke worries about the way information had been obtained and how inferences had been drawn from it. *Trigger questions* (e.g. "Why?" and "How?") are their creative cousins. They tend to initiate new thinking in certain directions. We will use the term "trigger question" for both roles. Such questions can be very effective. Many times in our interviews when no thinking was taking place, some small prompt opened flood gates of thought.

Experienced statisticians working in a collaborative or consulting environment learn to generate trigger questions which elicit pertinent context information, ideas and explanatory inferences from clients. The success of the dialogue between statistician and client may depend upon the quality of the trigger questions. No one taught our statistical consultants to ask the questions they do. Our consultants' statistics education had relied on the process: stimulus + experience + disposition $\longrightarrow$ pertinent trigger questions $\longrightarrow$ gaining critical ideas and knowledge about the context. This completely unstructured approach puts an enormous premium on experience. If statistical thinking is something that we teach rather than something simply to be absorbed by osmosis, then we have to give it structure. Structure can stimulate thinking, prevent crucial areas from being overlooked, and provide something to fall back on when you hit a brick wall.

The idea from the quality armory that we have found most powerful is the simple idea of intercon-

nected processes, with associated process diagrams, as a framework for analysing problems. It gives a structured way of breaking down activities, or mental processes, into components. It emphasises principal steps and their interrelationships and hides detail which can be uncovered subsequently when analysed as (sub)processes. Joiner's 7-Step Process (see Joiner 1994) is a thinking tool for working through for a quality improvement project. It leads the thinker through the steps of a project in time order and, within each step, gives lists of questions to prompt the crucial types of thinking that should occur there. The above approach is not new. Polya (1945) used it in *How to Solve It*, the most famous of all works on mathematical problem solving. In the quality arena, we see many serious attempts to capture essential elements of expert experience through creating thinking models/tools which can be used to approach specific types of problem. Underlying all of the above are two simple principles, which we have combined to form *systemise what you can, stimulate what you cannot.*

Schoenfeld (1987) distinguishes between a *description* which characterises a procedure and a *prescription* which characterises a procedure in sufficient detail to serve as a guide for implementing the strategy. PPDAC is a high-level description of a systematic approach to investigation. It identifies major elements. It can be used as the foundation for something that is much more of a prescription. This is a huge undertaking so what is presented here is merely indicative. The principles involved in our model fragments (Figs. 9 and 10) are:

- Systemise what you can, stimulate what you cannot.
- Use trigger questions to do the stimulation.
- Work from overviews and zoom in for the next level of detail.
- Keep the number of steps in any view of the process small to emphasise the most important relationships.

At any level, one drills down for more detail by clicking on a node in the process diagram (e.g. in an internet-type application). The area we have applied this to (in Figs. 9 and 10), is drilling down into the "Plan" node of PPDAC and then further down again into the "Measurement" node of the model of "Plan". We stopped at this level of detail and used sets of trigger questions about measurement (derived from the interviews) which are very general. Context-matter disciplines have built up enormous amounts of expertise about how to measure the things that are of great importance for research in their discipline. We have simply pointed to that with our questions. Models targeted at a particular application area could build in much more of that local expertise. Still, a general model such as ours could be useful for someone doing applied research in a less developed area, and for building in statistics students a more holistic feel for statistical investigation and the broad issues that need to be addressed. It prepackages some of the "strategic thinking" of breaking major tasks down into subtasks.

An attractive model element that we have not incorporated here, though it might be useful to do so, are lists of the tools that are helpful at particular nodes of a process. For examples, see Hoerl & Snee (1995). Process analysis tools provide us with a means of building up new bodies of "statistical theory" addressing critically important areas of the statistical process that statistics teachers are currently rather silent about. The results will be oversimplifications and sometimes gross oversimplifications, but then so are all our mathematical models. The theories should give students two opportunities to learn about and make sense of the statistical endeavour. First, the theory provides a scaffolding to use in forming a picture of some very complex processes. Second, once such a picture has been established, a more sophisticated understanding can be gained by considering ways in which the models are inadequate.

We conclude this subsection with a story related to us by some project students in a quality improvement course that sheds light on the complementary roles of theory and experience. The students first learnt some theory about quality improvement (including the role of statistical tools) via lectures and readings and found it all rather abstract and meaningless. On their first practical

project they floundered. The theory did not seem to help. But from those experiences the theory started to make sense. And by the second project it had started to work for them—its value had become obvious.

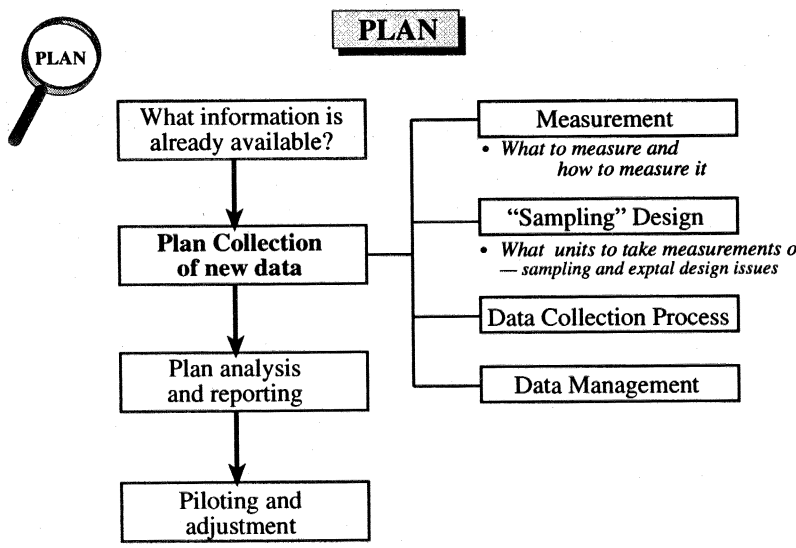

Figure 9. *Drilling down into the "Plan" node of PPDAC*

5 Discussion

The ultimate aim of statistical investigation is learning in the context domain of a real problem. Learning consists of the construction of mental models of the system under study. Statistics is itself a collection of abstract models ("models" is used in a very broad sense) which permit an efficient implementation of the use of archetypes as a method of problem solution. One abstracts pertinent elements of the problem context that map onto a relevant archetypical problem type, uses what has been worked out about solving such problems, and maps the answers back to context domain. There is a continual shuttling between the two domains and it is in this shuttling or interplay, that statistical thinking takes place—where the statistical rubber meets the real-world road. When it works, we gain real traction. Our abstraction processes bring clarity to thinking and efficiency to problem solving. However, when we use archetypes to solve problems, an enormous amount rides on the ability to do the mappings. And this is where the wheels so often fall off. Statistics education spends little time on developing the mappings. We must take more cognisance of the fact that the getting from the first stirrings of a practical problem to something like $y = \beta^T x + \varepsilon$, the point at which the theory of analysis typically kicks in, does not involve driving blithely across some small crack in the road, but rather it involves the perilous crossing of a yawning chasm down which countless investigations and analyses plummet to be lost without trace.

For successful problem solving, statistical thinking is not a separable entity. There is only holistic thinking that can and should be informed by statistical elements. The more relevant knowledge one

MEASUREMENT

(Including Classification)

Identify Key Characteristics

What features of the system are you interested in?
For each feature:
- What ideas about this feature are you trying to capture?
- Can you substitute something more specific?
- How well does this capture the idea?
- Is the idea you wish to capture by measurement inherently multidimensional in important ways or should a single measurement do?

Do these "key characteristics" adequately capture the essence of the real system?

Decide how to measure them

Experience in the field

Is there a generally accepted way of measuring this charactersitic in the field?
Is this currently accepted in the field as the best way to measure this entity?
If there is an accepted method and you want to use something else:
- What is your justification?

Have others tried to measure this? How did they do it?
- Are there known problems with their measure?

A fall-back

Can I draw on the experience of others in measuring similar charactersitics?

Anticipate problems

Validity & reliability

To what extent does this measurement really capture the characteristic I want to measure?
What are the practical implications of the extent to which it fails?
Will repeat measurements on the same units give very similar results?
Will different people making such measurements obtain very similar results?
Will different measuring instruments give very similar results?
If not, what impact will this have on the usefulness of any conclusions?

Analysis

Will I be able to analyse data containing measurements like this?
Will this measure make the analysis unnecessarily difficult?
Will another choice confer greater statistical efficiency?

"Audience" reaction

Will others be able to understand this measure?
Will the audience for the results accept that this is a sensible way to measure this?
Will I be able to understand the results of an analysis based on these measures?
Will I be able to communicate the results of an analysis based on these measures?

Practical implementation

Can I implement these measures in practice on the scale needed for the study?
Is the equipment/personnel required for this measure available? affordable?
Is the measure unacceptably or unnecessarily difficult? expensive? invasive?
Are there cheaper/easier/less invasive alternatives that will serve almost as well?

People: Do these measures take account of the psychological, cultural and perceptual differences of the people to be measured?

Can I do better?

Figure 10. *Drilling down further into the "Measurement" node of "Plan"*

has and the better one can connect it, the better one can do. In many research environments, statistical thinking is like breathing—everyone does it all the time, seldom being aware that it is happening. Statistics, the discipline, should be teaching people to "breathe" more effectively. However, we are dealing with complex and sophisticated thinking processes. We cannot expect, and indeed should be highly suspicious of, what W. Edwards Deming called "*instant pudding solutions*".

In Section 2, we identified several dimensions of the statistical thinking used in empirical enquiry: (1) the *investigative cycle*; (2) *types* of thinking; (3) the *interrogative cycle*; and (4) *dispositions* (see Fig. 1). We further discussed factors constraining the effectiveness of the thinking. Much of what we have identified relates to general problem solving skills being applied in a statistical context. One might think that for such skills a general thinking skills course such as those developed by de Bono is all that is needed. According to Resnick (1987), however, there is no empirical evidence that even these general skills are transferred to specific subject areas. She believes (p. 35) that thinking processes should be embedded into the discipline itself because, "*it provides a natural knowledge base and environment in which to practice and develop higher order* (thinking) *skills as . . . one must reason about something . . .* (and) *. . . each discipline has characteristic ways of reasoning . . .*." To carry out this embedding, we need more research into how these broad thinking skills are specifically used in statistics.

Omnipresent variation was presented as providing an important raison d'être for statistical thinking. In Section 3, we took the observable reality of variation in the concrete world as a starting point and endeavoured to cut through many of the confusions surrounding such abstract notions as "random variation" and their application to practical problem solving.

In Section 4, we discussed some techniques people have used to improve thinking. It seems to us that the rest of statistics can only benefit by following the lead of our colleagues in process and organisational improvement and develop tools that help us to think about, and to think through, parts of the statistical process that we are presently rather silent about. We can develop other forms of statistical model, other forms of statistical theory to deal with these areas. We stress that thinking tools are not a substitute for experience with investigation and data. Probably their most important purpose is to help us understand our experience and extend the range of situations to which we can apply it. But they may also re-initiate thinking that has become stalled.

As for the usefulness of the models presented, we have subsequently used this framework to develop judgement criteria to help students interpret statistically based information such as in media reports. We can see multiple uses even for a very simple model like the interrogative cycle (Fig. 1(c)). It could be used: to monitor thinking during problem solving; to help students become aware of their own thinking; as a tool for evaluating student thinking; and as a reference point against which to check learning opportunities provided to students. Do they, at least collectively, provide good opportunities for the students to experience all of these modes?—It turns out that many of the tasks we gave students did not! Nor did they measure up in terms of types of thinking.

Can thinking tools work? The people in process and organisational improvement and Polya and his successors in mathematics believe so. Are they panaceas? There is nothing certain or cut-and-dried in applied statistics. The real world is a messy, complicated place. We are reminded of David Moore's distinction between mathematics and statistics, "Mathematical Theorems are true; statistical methods are sometimes useful when used with skill." We cannot expect more from our new tools than from our traditional ones. Statistics is not some hiking tent that can be erected in an afternoon. It is an enormous edifice. Most of the work in our writing and teaching, however, has gone into constructing its upper levels. But, with advancing technology inexorably shifting the balance of human statistical effort from processing and computation to thinking, we need to do some emergency work on the foundations to ensure that the whole structure stands steadily on the ground.

References

Bartholomew, D. (1995). What is Statistics? *J.R. Statist Soc. A*, **158**, Part 1, 1–20.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning and the search for causes: Do we need a probabilistic revolution after we have taught data analysis? In *Research Papers from The Fourth International Conference On Teaching Statistics, Marrakech 1994*, Ed. J. Garfield. MN: University of Minnesota.

Britz, G., Emerling, D., Hare, L., Hoerl, R. & Shade, J. (1997). How to Teach Others to Apply Statistical Thinking. *Quality Progress*, June 1997, 67–79.

Cox, D.R. (1992). Causality: Some Statistical Aspects. *J. R. Statist. Soc. A*, **155**, Part 2, 291–301.

Cox, D.R. (1993). Causality and Graphical Models. In *Bulletin of the International Statistical Institute, Proceedings of the 49th Session* Vol. 1, pp. 365–389. Voorburg: International Statistical Institute.

Cox, D.R. (1997). The current position of statistics: A personal view (with discussion). *International Statistical Review*, **65**, 261–290.

Dransfield, S.B., Fisher, N.I. & Vogel, N.J. (1999). Using statistics and statistical thinking to improve organisational performance. *International Statistical Review*, **67**, 99–150.

Falk, R. & Konold, C. (1992). The Psychology of Learning Probability.In *Statistics for the Twenty-First Century*, Eds. F. & S. Gordon, pp. 151–164. *MAA Notes*, Number 29. Washington, DC: The Mathematical Association of America.

Gail, M. (1996). Statistics in Action. *Journal of the American Statistal Association*, **91**, 1–13.

Gal, I., Ahlgren, C., Burrill, G., Landwehr, J., Rich, W. & Begg, A. (1995). Working group: Assessment of Interpretive Skills. *Writing Group Draft Summaries Conference on Assessment Issues in Statistics Education*, pp. 23–25. Philadelphia: University of Pennsylvania.

Garfield, J. & Ahlgren, A. (1988). Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education*, **19**, 44–63.

Hill, A.B. (1965). The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine*, **58**, 295–300.

Hoerl, R., Hahn, G. & Doganaksoy, N. (1997). Discussion: Let's Stop Squandering Our Most Strategic Weapon. *International Statistical Review*, **65**, 147–153.

Hoerl, R. & Snee, R.D. (1995). Redesigning the introductory statistics course. *CQPI Technical Report No. 130*, Center for Quality and Productivity Improvement, University of Wisconsin-Madison.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945–970.

Joiner, B. & Gaudard, M. (1990). Variation, Management, and W. Edwards Deming. *Quality Progress*, December, pp. 29–37.

Joiner, B. (1994). *Fourth Generation Management*. New York: McGraw-Hill Inc.

Konold, C. (1994). Understanding Probability and Statistics through Resampling. In *Proceedings of the First Scientific Meeting of the International Association for Statistical Education*, Eds. L. Brunelli & G. Cicchitelli, pp. 199–211. Perugia: University of Perugia.

MacKay, R.J. & Oldford, W. (1994). *Stat 231 Course Notes Fall 1994*. Waterloo: University of Waterloo.

Mallows, C. (1998). The zeroth problem (1997 Fisher Memorial Lecture). *The American Statistician*, **52**, 1–9.

Moore, D. (1990). Uncertainty. In *On the shoulders of giants: new approaches to numeracy*, Ed. L. Steen, pp. 95–137. Washington, DC: National Academy Press.

Moore, D. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review*, **65**, 123–165.

Pea, R. (1987). Cognitive Technologies for Mathematics Education. In *Cognitive Science and Mathematics Education*, Ed. A. Schoenfeld, pp. 89–122. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Pfannkuch, M. & Brown, C. (1996). Building on and Challenging Students Intuitions about Probability: Can We Improve Undergraduate Learning? *Journal of Statistics Education*, www.amstat.org/publications/jse/v4nl/pfannkuch.html.

Pfannkuch, M. (1996). Statistical Interpretation of Media Reports. In *New Zealand Statistical Association Research in the Learning of Statistics Conference Proceedings*, Eds. J. Neyland and M. Clark, pp. 67–76. Wellington: Victoria University.

Pfannkuch, M. (1997). Statistical Thinking: One Statistician's Perspective. In *People in Mathematics Education. Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia*, Eds. F. Biddulph and K. Carr, pp. 406–413. Hamilton: MERGA Inc.

Pfannkuch, M. & Wild, C. (1998). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. (unpublished manuscript).

Polya, G. (1945). *How to Solve It: A New Aspect of Mathematical Method*, Princeton: Princeton University Press.

Pyzdek, T. (1990). There's no such thing as a common cause. *ASQC Quality Congress Transactions–San Francisco*, pp. 102–108.

Resnick L. (1987). *Education and Learning to Think*. Washington DC: National Academy Press.

Resnick, L. (1989). Treating Mathematics as an Ill-structured discipline. In *The Teaching and Assessing of Mathematical Problem Solving, Volume 3*, Ed. R. Charles and E. Silver, pp. 32–60. Reston, VA: Lawrence Erlbaum Associates NCTM.

Schoenfeld, A. (1983). Episodes and Executive Decisions in Mathematical Problem-Solving In *Acquisition of Mathematics Concepts and Processes*, Ed. R. Lesh and M. Landau, pp. 345–395. New York: Academic Press.

Schoenfeld, A. (1987). Cognitive Science and Mathematics Education: An Overview. In *Cognitive Science and Mathematics Education*, Ed. A. Schoenfeld, pp. 1–31. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Shaughnessy, M. (1992). Research in Probability and Statistics: Reflections and Directions. In *Handbook of Research on Mathematics Teaching and Learning*, Ed. D. Grouws, pp. 465–494. New York: MacMillan.

Snee, R. (1990). Statistical Thinking and its Contribution to Quality. *The American Statistician*, **44**, 116–121.

Stewart, I. (1989). *Does God Play Dice?* London: Penguin Books.

Tversky, A. & Gilovich, T. (1989). The "hot hand": Statistical reality or cognitive illusion? *Chance*, **2**, 31–34.
Wild, C.J. (1994). On Embracing the "Wider View" of Statistics. *The American Statistician*, **48**, 163–171.

Résumé

Le présent article concerne les processus mentaux impliqués dans la pensée statistique prise dans un sens large, depuis la formulation de problèmes jusqu'á leur solution. Il tire ses sources de la littérature sur le sujet ainsi que d'entrevues auprès d'étudiants et de praticiens en statistique, conçues pour identifier leurs processus de raisonnement statistique. De ces entrevues, nous avons identifié un cadre conceptuel quadridimensionel applicable à la pensée statistique dans le domaine de la recherche empirique. Ce cadre est composé d'un cycle d'investigation, d'un cycle d'interrogation, de types de pensée et de dispositions. Nous avons amorcé la caractérisation de ces processus par des modèles pouvant servir de base à la création d'outils ou cadres intellectuels aidant la résolution de problèmes. Des outils de ce types pourraient complémenter les modèles mathématiques déjà utilisés en analyse en plus de couvrir certains aspects de la recherche statistique que les modèles mathématiques ne peuvent pas satisfaire, particulièrement les aspects associés à la synthèse des types contextuel et statistique de compréhension. L'élément central apparaissant dans les définitions de la pensée statistique ayant fait l'objet de publication est celui de la "variation". Nous discutons aussi le rôle de la variation dans l'approche statistique de problèmes pratiques, incluant la recherche de causes.

Discussion

T.M.F. Smith

University of Southampton, UK

The topic of this thoughtful paper is the modes of thinking that underlie efficient empirical enquiries. The authors, henceforth WP, title the paper statistical thinking but, in fact, the topic is much wider and embraces the thinking and learning processes that form the basis of all scientific investigations. If throughout the paper you replace the word statistical by scientific, or more generally systematic, then very little would have to be changed. The models represented by Fig. 1–10 remain valid and provide an excellent framework for discussing scientific thinking. The only term that is missing is creativity, the mode of thinking that leads to the greatest advances in science. But creativity is impossible to teach and WP are concerned with processes that can, and should, be taught, and that form the basis of all scientific enquiry from the most humdrum to the most creative.

A key word that is repeated throughout the paper is context. All thinking, including statistical thinking, takes place within a context which needs to be understood by all who are conducting the enquiry. WP make a strong case for multi-disciplinary teams who challenge assumptions, bring different backgrounds to enquiries, and generate the synthesis of ideas that lead to understanding and future progress. WP argue that statisticians should frequently be members of these teams, but they include examples where conventional statistical methods, as currently taught, do not add much value to the enquiry. To justify team membership, statisticians must be seen to add value that cannot be added by scientists from other disciplines. In the examples from both students and consultants this is not always the case. In discussing context we learn that "Of all our statisticians, however, only the one operating in quality improvement seemed to use tools (e.g. cause and effect diagrams and process mapping techniques) to analyse the nature of the problem itself." How many statisticians have been taught to use these tools? Later on, in the discussion of causation, we find the statistician as Cassandra, the harbinger of doom, warning that correlation does not imply causation and that results may be meaningless if they could have arisen by chance. Statistical thinking can be very negative at times and can stifle imaginative enquiry. Is that really our main contribution? Sometimes yes, since there is a danger in jumping too quickly to conclusions in areas such as medicine. But this danger may not exist in other areas and statisticians must be aware of the context in which they are operating