STATISTICAL COGNITION: TOWARDS EVIDENCE-BASED PRACTICE IN STATISTICS AND STATISTICS EDUCATION

RUTH BEYTH-MAROM
Department of Education and Psychology, The Open University, Israel
ruthbm@openu.ac.il

FIONA FIDLER
School of Psychological Science, La Trobe University, Melbourne, Australia
f.fidler@latrobe.edu.au

GEOFF CUMMING
School of Psychological Science, La Trobe University, Melbourne, Australia
g.cumming@latrobe.edu.au

ABSTRACT

Practitioners and teachers should be able to justify their chosen techniques by taking into account research results: This is evidence-based practice (EBP). We argue that, specifically, statistical practice and statistics education should be guided by evidence, and we propose statistical cognition (SC) as an integration of theory, research, and application to support EBP. SC is an interdisciplinary research field, and a way of thinking. We identify three facets of SC—normative, descriptive, and prescriptive—and discuss their mutual influences. Unfortunately, the three components are studied by somewhat separate groups of scholars, who publish in different journals. These separations impede the implementation of EBP. SC, however, integrates the facets and provides a basis for EBP in statistical practice and education.

Keywords: *Statistics education research; Statistical cognition; Statistical reasoning*

1. BACKGROUND

A wide range of research is relevant for improving statistical practice and statistics education, but we worry that this research is too fragmented for most effective use. We identify three facets of this research, and propose that the concept of *statistical cognition* can help bring these together, and provide a stronger basis for evidence-based practice (EBP) in statistics and statistics education. As an introductory example, consider confidence intervals (CIs), and three lines of discussion involving them.

First, for almost a century, mathematical statisticians have been studying CIs—developing theory and new applications, investigating robustness, and making comparisons with other inferential techniques. Second, within statistics, and in research fields that use statistics, there has been some discussion about possible misunderstandings of CIs; textbook authors also consider how to explain CIs, and possible misconceptions. However there has been almost no empirical study of how students and researchers think about CIs, or about misconceptions they may have. Third, there have been persistent calls for much wider use of CIs, in preference to null hypothesis significance testing (NHST), in psychology and other disciplines (e.g., Wilkinson et al., 1999). Reformers have

claimed CIs lead to better research decision making than NHST, and that students can more easily and successfully learn about CIs than NHST (e.g., Schmidt & Hunter, 1997). Our worry is that the evidence base especially for the second and third lines of discussion is sadly deficient, and that the three lines are not sufficiently integrated.

The first discussion, or facet, we referred to above was *normative*: theory of CIs, and techniques for their application, developed within mathematical statistics. The second considered how researchers, students, or others think about CIs—their informal statistical reasoning. This is the *descriptive* facet, which focuses on the cognition of using or teaching statistics. The third was the recommendation to replace NHST with CIs, and this is obviously *prescriptive*. The prescriptive facet seeks to improve statistical practice, and statistics learning. It might, for example, provide evidence about which CI diagrams and explanations are most effective in helping students achieve correct conceptualisations, as well as about which graphical designs and CI interpretations most successfully communicate research results.

By the 1980s the distinction between normative, descriptive and prescriptive was commonplace in judgment and decision making literature. "Decision Making: Descriptive, Normative and Prescriptive Interactions" was the name of a conference held in Boston at the Harvard Business School in 1983, the product of which was an edited book with this title (Bell, Raiffa, & Tversky, 1988). Those authors suggested the following taxonomy:

Descriptive: (1) Decisions people make; (2) How people decide.
Normative: (1) Logically consistent decision procedures; (2) How people should decide.
Prescriptive: (1) How to help people to make good decisions; (2) How to train people to make better decisions. (p. 1-2)

For the purpose of the current discussion, we could substitute "statistical inferences" for "decisions." We are interested in the mutual influences and contributions of these three facets, as well as their integration. One motivation for integration is to provide a more cohesive and complete evidence base for statistical practice and education.

Evidence-based practice (EBP) has a long history in medical decision making. The Institute of Medicine (2001) defined EBP as "the integration of best research evidence with clinical expertise and patient values" (p. 147). Psychology, nursing, social work, and other professional disciplines are progressively advocating and adopting EBP (Trinder & Reynolds, 2000). *Evidence-Based Medicine*, *Evidence-Based Child Health*, *Evidence-Based Communication Assessment and Intervention*, *Evidence-Based Complementary and Alternative Medicine*, *Evidence-Based Library and Information Practice* are all relatively new journals aimed to alert professionals to important theoretical and empirical advances in their profession that might contribute to improved decision making in their professional practice. Similarly, a desire to ensure that students meet high standards has increased the demand for EBP in education (Davies, 1999). Statisticians and statistics educators should likewise adopt EBP by, wherever possible, using relevant evidence from research to guide what they do.

Within medical EBP, successful implementation of research into practice requires integration of three core elements: relevant *evidence*, the context or *environment* into which the research is to be placed, and the *method* or way in which the process is accomplished (Kitson, Harvey, & McCormack, 1998). There is some correspondence between these elements and our normative, descriptive and prescriptive facets, respectively. If a statistician is advising a researcher about data analysis for a report, normative information about statistics provides the *evidence*, for example, statistical theory about correlation. Descriptive information about likely misunderstandings of

correlation by the readers of a journal article is part of the researcher's *context*; and prescriptive information—if available—suggests how most effectively to present correlations and thus provides a *method*.

We therefore believe that these three lines of research are necessary to build an evidence base for statistical practice and education, and that adoption of EBP directly depends on the integration of these fragmented research facets. In this article, we first introduce statistical cognition—as a concept and an integrative field—in further detail. Second, we explore the interactions among the normative, descriptive, and prescriptive facets. Some of these interactions may be obvious, but others are subtle and still others virtually missing. Exploring these relationships helps identify gaps in current research, and priorities for future research. Third, we describe two examples to illustrate these interactions in statistics teaching and practice. We then briefly examine institutional and sociological factors that have contributed to the fragmentation of the normative, descriptive, and prescriptive facets of research. Finally, we explore how statistical cognition may overcome some of the barriers that currently impede integration, and make recommendations about how this integrated field should proceed.

2. STATISTICAL COGNITION

Cognition is usually defined as the mental processes, representations, and activities involved in the acquisition and use of knowledge. Statistical cognition is accordingly defined as the processes, representations, and activities involved in acquiring and using *statistical* knowledge. What are the issues relevant in the study of statistical cognition? One aspect is how people acquire and use statistical knowledge and how they think about statistical concepts—this is the descriptive facet of statistical cognition. The study of how people *should* think about statistical concepts—the normative—is also an important aspect of statistical cognition as this is often what we are exposed to (e.g., in school) and it is also the standard to which our performance is usually compared. Finally, the question of closing the gap between the descriptive (the "is") and the normative (the "should")—the prescriptive—is a critical issue in statistical cognition.

As such, statistical cognition is a field of theory research and application concerned with normative, descriptive, and prescriptive aspects. It focuses on (a) developing and refining normative theories of statistics and their application, (b) developing and testing theories explaining human thinking about and judgment in statistical tasks, and (c) developing and testing pedagogical tools and ways of communication for the benefit of practitioners and teachers.

Statistical reasoning, a term already widely used (Garfield, 2002; Garfield & Gal, 1999), concerns the mental processes which shape the process and representations of statistical cognition. As such, it is concerned mainly with the descriptive facet. However, statistical cognition, like mathematical cognition, takes a broader approach encompassing normative and prescriptive research, in addition to the descriptive research found in the literature on statistical reasoning and in the experimental and educational psychology literatures. Statistical cognition therefore integrates the three lines of research we believe are needed for effective EBP.

3. THE THREE FACETS OF STATISTICAL COGNITION: NORMATIVE, DESCRIPTIVE AND PRESCRIPTIVE

The science of statistics contributes most to the normative facet of statistical cognition. It includes simple rules (e.g., the conjunction rule of probability), theorems and laws (e.g., Bayes' theorem, the law of large numbers), as well as models (e.g., for

estimation and inference). Statisticians may agree that a particular normative solution to a problem is best, or may hold differing views as to which normative model should be applied. Both Frequentist and Bayesian approaches have been developed and advocated within the normative facet, as has theory for both NHST and estimation. Whether consensus or controversy dominates, such rules, models, and approaches comprise the normative facet of statistical cognition.

Dissemination of statistical information (beginning in the 19th century) about many aspects of society has increased the need for laypersons as well as professionals to understand statistical concepts. Many reports in the mass media (about psychological, medical, economic, or political issues) can only be correctly comprehended with an understanding of statistics. As early as the beginning of the 20th century, H. G. Wells emphasised the importance of teaching statistical reasoning to produce an educated citizen, with statistical reasoning being as important as reading and writing (Huff, 1973, p. 6). In the early 1980s the concept of 'statistical numeracy' was first introduced as a sub-category of 'numeracy':

> Statistical numeracy requires a feel for numbers, and appreciation of appropriate levels of accuracy, the making of sensible estimates, a commonsense approach to the use of data in supporting an argument, the awareness of variety of interpretation of figures, and a judicious understanding of widely used concepts such as means and percentages. (Cockcroft, 1982, paragraph 781)

The broader term 'statistical literacy' (Ben-Zvi & Garfield, 2004; Gal, 2002; Wallman, 1993) later replaced statistical numeracy, and became an important goal yet to be achieved. The need to enhance statistical literacy has been gradually recognised with the publication of psychological research assessing intuitive statistical reasoning (e.g., Edwards, 1968; Meehl, 1954; Tversky & Kahneman, 1974) and studying the cognitive processes involved (e.g., Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982; Sedlmeier, 1999). These developments shaped two lines of theory, research, and applications: the descriptive and the prescriptive approaches.

"Man as an intuitive statistician" (Peterson & Beach, 1967) was the first comprehensive publication on intuitive statistical reasoning and it opened a long-lasting debate about lay persons' as well as experts' capabilities. Tversky and Kahneman's (1974) seminal work replaced Peterson and Beach's optimistic view with the heuristic and biases model: Intuitive statistical judgments are often based on a limited number of simplifying heuristics rather than on more formal and extensive algorithmic processing. These heuristics can give rise to systematic errors, or biases. These lines of research—the evaluation of people's statistical reasoning and the cognitive processes underlying them—are the core of the descriptive aspect of statistical cognition.

Statistical education aims to improve statistical reasoning. The best approaches and tools for reaching this goal, and the pedagogical prescriptions for the teaching of statistics, should be based on the art, science, and profession of teaching. Learning by doing (e.g., Glaser & Bassok, 1989; Smith, 1998), authentic learning (e.g., Donovan, Bransford, & Pellegrino, 1999; Mehlinger, 1995) and situated cognition (e.g., Brown, Collins, & Duguid, 1989) are examples of educational or instructional theories that have direct pedagogical recommendations.

A statistical consultant may advise that a particular model and statistical analysis is appropriate for the data of interest—relying on normative considerations. The question then becomes how the results will be written up for publication, and that is a question of statistical communication: What numerical, graphical or other information should be presented so that target readers will understand most accurately what was found and what conclusions are justified? Those questions should be in the forefront of the mind of the statistical consultant, as well as the researcher, and it is the job—we would argue—of

statistical cognition to provide research-based guidance as to how statistical communication can be best accomplished. Similarly, the discipline of statistics provides much content for the statistics curriculum, but it is the job of statistical cognition to provide guidance for teachers on how to best achieve accurate and appropriate statistical learning.

Each of the three approaches has theory, research, and applications rooted primarily in different disciplines (statistics, psychology, education). As we have indicated, we believe there has been insufficient interaction between them. We hope that statistical cognition can encourage closer collaboration among the approaches, and thus develop a body of research that can support EBP in statistics. This body of research should focus on projects like statistical reasoning of laypersons as well as experts; developmental aspect of statistical reasoning along the life span; cognitive, social and neurological processes that underlie statistical reasoning; and testing the efficiency of instructional techniques, approaches, and tools.

Figure 1 illustrates in a schematic way the three facets of statistical cognition, and the arrows indicate paths of influence. The normative facet (N) specifies what statistical techniques can correctly be applied in a given situation; it is potentially informed by the full body of knowledge that is mathematical statistics. The descriptive facet (D) comprises knowledge of how people think about statistical concepts, what messages they receive when inspecting a statistical presentation, and their statistical misconceptions and biases. Psychology has provided most of the information in D, yet this information is scanty and there are many important gaps that need further research. The prescriptive facet (P) comprises knowledge about how to achieve successful statistical communication and education. This knowledge, such as it is, has largely come from psychology and education, and again much additional knowledge is needed in this facet. The contribution of the normative facet (N) to the prescriptive (P) is large and probably straightforward to grasp: It is probably most natural and common to base advice or teaching on statistical theory. There can perhaps (the dotted arrow) be influence in the reverse direction, when experience with advising or teaching (that's P) prompts development of additional theory (N). The next sections will focus on the two-way influences between N and D, and between D and P. We consider both the known and the potential contributions relevant to each arrow.

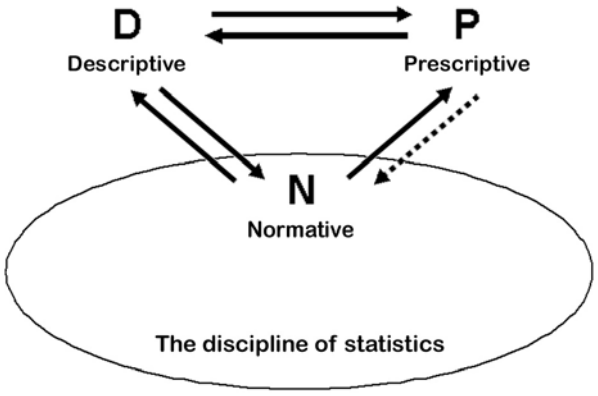

Figure 1. Schematic relations between the proposed three facets of statistical cognition

4. CONTRIBUTION OF THE NORMATIVE TO THE DESCRIPTIVE

The normative rules, theories, and models of the science of statistics are the standards recommended for summarizing data, interpreting it, and evaluating hypotheses. These are the norms used by professionals when analysing empirical research or advising researchers. However, these norms have also been used as standards to which intuitive statistical reasoning (of laypeople and experts) is compared. For example, people's performance in solving conjunction tasks has been compared to the predictions of the conjunction rule: $P(A\&B) \leq P(A)$ and $P(A\&B) \leq P(B)$ (Tversky & Kahneman, 1983). Normative standards have been used similarly in research on people's judgments of disjunctive probabilities (Bar-Hillel & Neter, 1993), conditional probabilities (Pollatsek, Well, Konold, Hardiman, & Cobb, 1987), effects of sample size (Bar-Hillel, 1979), judgment of randomness (Falk & Konold, 1994, 1997), interpreting p-values in hypothesis testing (Falk, 1986; Oakes, 1986)—to mention but a few cases.

A normative model can thus provide a theoretical framework for describing how people should perform a task. It can also identify a set of logically possible deviations from the model, which can be tested empirically. Such an approach was used by Fischhoff and Beyth-Marom (1983). They adopted Bayesian inference as a general framework for characterizing people's hypothesis evaluation behaviour in terms of its consistency with or departures from the model. They identified the kinds of systematic deviations from the Bayesian model that could, in principle, be observed, and presented evidence demonstrating their actual existence. Normative models provide a reference for the evaluation of people's performance in statistical tasks (the descriptive facet).

The choice of the appropriate normative model may seem obvious, but sometimes is debatable, or may be thrown into doubt after further consideration of descriptive results. Gigerenzer (1991), for example, argued that probability theory is imposed as a norm for judgments about a single event in research on the conjunction fallacy, and this would be considered misguided by statisticians who hold that probability theory is about repeated events. A further example is Cohen's (1979) questioning of the choice of a Bayesian model as a normative standard in Tversky and Kahneman's (1974) descriptive work; he suggested an alternative normative Baconian model. Thus, the choice of a normative standard to which people's performance is compared must be made with much care, being sure that the assumptions underlying the normative model (e.g., random sampling), are also part of the judgmental task performed by people.

5. CONTRIBUTION OF THE DESCRIPTIVE TO THE NORMATIVE

How can judgmental tasks, and people's performance of them, contribute to the relevant normative model? The historical account of NHST, empirical research on the understanding of p-values and, more generally, of people's intuitive inferential reasoning, provides one example of such a contribution.

NHST in its contemporary form (a hybrid of two schools of thought, one associated with Fisher, the other with Neyman and Pearson) was gradually applied in empirical research from 1940 (Hubbard & Ryan, 2000). There has been controversy about NHST since its inception, and the number of published works critical of it has increased dramatically since then (Anderson, Burnham, & Thompson, 2000).

The most common arguments against NHST refer to a catalogue of misconceptions about p-values. This catalogue (which is descriptive) has been built over many years from teachers' observations (e.g., Schmidt & Hunter, 1997), surveys of journal reporting practices (e.g., Finch, Cumming, and Thomason, 2001; Fidler et al., 2005) and empirical studies with researchers and students (Haller & Krauss, 2002; Kalinowski, Fidler, &

Cumming, 2008; Oakes, 1986). That the misconceptions are widespread and robust is well known and often demonstrated. They have also been compiled and summarised often. Kline (2004), for example, listed five common fallacies in the interpretation of p-values and eight common fallacies in reaching conclusions after deciding to reject or failing to reject the null hypothesis based on a p-value.

There are certainly advocates of statistical reform who believe that such misconceptions are the overwhelming, if not sole, problem with NHST. Rossi (1997), for example, stated "whereas some see significance testing as inherently flawed, I believe the problem is better characterised as the misuse of significance testing" (p. 175). However, there are others who hold the position that, even if used and interpreted properly, NHST contributes little knowledge and "is not the way any [proper] science is done" (Cohen, 1994, p. 999). A stronger expression of this position is that the procedure is itself fundamentally flawed; that NHST has a "flawed logical structure" (Falk & Greenbaum, 1995, p. 75).

There is also a third position, which draws the previous two together, and illustrates how the descriptive can contribute to the normative. This is the position that NHST is so widely misinterpreted *precisely because* the underlying logic is flawed. As Kline (2004) explained, "false beliefs may not be solely the fault of the users of statistical tests. … This is because the logical underpinnings of contemporary NHST are not entirely consistent" (p. 9). Kline is referring to the conflicting Fisherian and Neyman-Pearsonion paradigms that have become the institutionalised hybrid of NHST. Schmidt and Hunter (1997) provided another illustration of how descriptive considerations have challenged the normative status of NHST: "Any teacher of statistics knows that it is much easier for students to understand point estimates and CIs than significance testing with its strangely inverted logic" (p. 56). For these critics, challenges to the normative status of NHST have (at least in part) emerged from descriptive work on misconceptions.

Another alternative to NHST is Bayesian hypothesis testing, which differs from NHST in its interpretation of probability, and on these three principles: (1) Prior probabilities have to be taken into account; (2) alternative hypotheses play a role in the testing of a null hypothesis; and (3) the focus of analysis is P(H|D), and not P(D|H), where H is a hypothesis and D some data. We believe Bayesian methods have not been widely accepted at least in part because of users' misconceptions. That is, that their normative status has been in part determined by obstacles that are descriptive in nature. Research on the base rate fallacy (Bar-Hillel, 1980) demonstrated how people tend to ignore base-rates, thus behaving like null hypothesis statistical testers. Research on pseudo-diagnosticity (Beyth-Marom, 1990; Beyth-Marom & Fischhoff, 1983) indicated that people often base their updating of a hypothesis on the magnitude of P(D|H), ignoring P(D|~H), thus again behaving like null hypothesis statistical testers. There is also overwhelming evidence that people often confuse P(H|D) and P(D|H) and use, incorrectly, P(D|H) as their estimate of P(H|D) (Eddy, 1982; Haller & Krauss, 2002; Oakes, 1986). Thus, although people demonstrate severe misconceptions of the NHST model, by ignoring base rates and the relevance of alternative hypotheses, and by using P(D|H) for P(H|D), their intuitions remain more in line with the NHST model than the Bayesian one.

Descriptive findings also shed light on the history and development of normative models in science more broadly. Research on the perception of different scientific concepts (e.g., in physics, mathematics, and biology) by laypersons and experts has repeatedly shown that intuitive concepts deviate systematically from normative ones. Often the intuitive beliefs were similar to earlier, and now discredited, scientific theories. Erickson (1980), for example, investigated the change of children's viewpoints about heat

from a Caloric viewpoint to a Kinetic one. This developmental change has its counterpart in the evolution of physics. Perhaps a parallel in statistical inference is yet to occur: Researchers' intuitions, and their statistical practices, are both still largely at the NHST stage; advancement to, for example, Bayesian thinking, and Bayesian techniques, is still largely for the future. (We recognise that development of Bayesian methods pre-dated development of NHST; it is widespread adoption that is the focus here.)

Naïve statistical concepts can thus influence the normative theories of statistics; first, by offering insight into their evolution and, second, by questioning their validity and contributing to their development and change.

6. CONTRIBUTION OF THE DESCRIPTIVE TO THE PRESCRIPTIVE

The idea that descriptive should influence prescriptive may be familiar, but we believe that fully exploiting the potential contribution of the descriptive facet to the prescriptive is the most serious challenge in the triangle of Figure 1. Many practitioners and teachers of statistics, and authors of statistical textbooks, are only vaguely aware of the substantial cognitive literature on statistical reasoning and the contributions it can make.

Students young or old don't enter the learning arena 'tabula rasa' (Pinker, 2002), but already holding beliefs about scientific concepts and processes. They also have everyday meanings for words that are used in a more specialized way in science. These beliefs might help or hinder learning depending on their consistency or discrepancy with what is taught.

From this perspective, educators have been interested in students' 'preconceptions', 'naïve conceptions', or 'naïve theories'. If those were found to be inconsistent with formal concepts to be taught they were regarded as 'misconceptions' or 'alternative conceptions'. Misconceptions may come from strong word association, confusion, conflict, or lack of knowledge (Fisher, 1985). They usually share the following characteristics: (a) they are at variance with normative conceptions held by experts in the field; (b) they tend to be pervasive (shared by many different individuals), and (c) they are often highly resistant to change, at least by traditional teaching methods. Thus, special teaching methods have to be developed. For example, some educators recommend that teachers should be given numerous examples of how to identify misconceptions held by pupils and strategies to change them (Lawrenz, 1986; Smith & Anderson, 1984). Others have suggested starting the teaching with students' ideas and then devising teaching strategies to take some account of them (Engel Clough & Wood-Robinson, 1985).

Most research on naïve conceptions and misconceptions—the descriptive element of statistical cognition—originated in cognitive psychology: how people reason under uncertainty. As uncertainty and statistical information surrounds us, efficiently coping with it is essential for our everyday conduct. Moreover, statistical reasoning is a tool used by experts in carrying out and interpreting research. Thus teaching of (normative) statistics is essential in schools—for the developing of good statistical reasoning—and in university—for doing research and interpreting its results. However, there is evidence (e.g., Abelson, 1995; Sedlmeier & Gigerenzer, 1989), as well as widespread classroom experience, to suggest that the teaching of statistics is often not very successful. In a literature review of the teaching of statistical reasoning to students at college and precollege levels, Garfield and Ahlgren (1988) concluded, two decades ago, that "little seems to be known about how to teach probability and statistics effectively" (p. 45). More recent research has had some impact on college teaching, but many courses remain unaffected by its outcomes (Garfield, Hogg, Schau, & Whittinghill, 2002; Ben-Zvi & Garfield, 2004). In other sciences, at least some educators are aware of the relevance of

misconceptions (the descriptive facet) to the effective teaching of science (the prescriptive facet), but it seems that statistical instructors are less often aware of students' statistical misconceptions, and have few instructional tools designed to overcome them. Cognitive psychology, however, can offer considerable research on naïve conceptions and misconceptions, and how people reason under uncertainty—the descriptive element of statistical cognition.

Descriptive research on informal statistical reasoning might contribute to statistics teaching not only by identifying misconceptions, but also by describing the processes underlying them. This research can guide the development of effective teaching strategies. In his book *Improving statistical reasoning: Theoretical models and practical implications* Sedlmeier (1999) identified four descriptive explanatory models of statistical reasoning, and derived from them implications for statistical teaching. According to Sedlmeier's 'adaptive algorithms' explanation, the human mind is equipped with evolutionarily acquired cognitive algorithms that are able to solve complicated statistical tasks. These algorithms work for frequencies, but not for probabilities or percentages. The instructional implication is that to improve performance we should teach people how to translate from the format given in the task (e.g., probabilities) into natural frequencies (Gigerenzer & Hoffrage, 1995).

The heuristics and biases cognitive psychology literature has adopted dual-process theory (Sloman, 1996; Stanovich & West, 2000), which identifies two quite different cognitive modes, System 1 (S1) and System 2 (S2), approximately corresponding to the common sense notions of intuitive and analytical thinking. The two systems differ in various ways, most notably on the dimension of accessibility: how fast and how easily things come to mind. Many of the non-normative answers people give to statistical (as well as other) questions can be explained by the quick and automatic responses of S1, and the frequent failure of S2 to intervene in its role as critic of S1. Based on this dual-process theory, Kahneman in his Nobel Prize lecture set an agenda for research:

> To understand judgment and choice we must study the determinants of high accessibility, the conditions under which S2 will override or correct S1, and the rules of these corrective operations. Much is known of each of the three questions, all of which are highly relevant to the teaching of statistical reasoning. (Kahneman, 2003, p. 716)

Leron and Hazzan (2006) demonstrated how dual-process theory and empirical results from heuristic and biases research might shed light on mathematics education. They argued that the most important educational implication is "to train people to be aware of the way S1 and S2 operate, and to include this awareness in their problem solving toolbox" (p. 123). Such a toolbox is relevant also for statistical reasoning.

Communication of statistical information in newspapers and magazines, as well as in statistical textbooks and courses, includes many words used for statistical concepts that are also used in common language, such as 'or', 'chance', 'randomness', 'confidence', 'precision', and 'correlation'. Often, the meaning in everyday language is similar to the technical meaning in the science of statistics. However, sometimes the two concepts do not overlap. For example, Beyth-Marom (1982) showed how laypersons interpret correlation between two asymmetric variables, such as 'pneumonia: pneumonia vs. no pneumonia', where the values have differential status, and the variable has the name of one of its values. Participants interpret such correlations as the tendency of the two 'present' values to coexist, thus interpreting relationship between variables as a relationship between values, and ignoring all other information relevant for the evaluation of statistical correlation. This everyday interpretation of correlation is consistent with the *Oxford English Dictionary* (2007) definition of correlation as "a mutual relationship." This dictionary, as well as *The American Heritage* (2000) and *Webster's Online* (2007)

dictionaries, present two or more definitions of correlation: one specific for statistics (mentioning variables) and the other the common everyday interpretation (mentioning entities or things). Falk and Konold (1994, 1997) showed a similar phenomenon in the perception of randomness.

When ordinary language is used for reasoning, conceptions and misconceptions are often shaped by the nature of the social interaction and the conversation taking place (Grice, 1975). "Respondents [students] deviate from the judgments predicted by the normative model considered relevant by the experimenter [teacher] by using rules of conversational inference very different than those assumed by the experimenter [teacher]" (Hilton, 1995, p. 266). Recognition of linguistic and conversational factors, as well as being alert to possible discrepancies between statistical and conversational meanings, is likely to have practical pedagogical implications for improving statistical understanding.

A further contribution of D to P is developmental research on statistical reasoning, beginning with Piaget and Inhelder (1975). The introduction of statistics into the school curriculum has prompted more attention to understanding developmental aspects of statistical literacy. A number of models of cognitive development in probability and statistics have been proposed (Biggs & Collis, 1991; Jones, Langrall, Thornton, & Mogill, 1997; Mooney, 2002; Watson & Callingham, 2003). According to these, instructional materials should be appropriate for students' age and cognitive development. For example, the model suggested by Jones and his colleagues (1997, 1999) includes four statistical concepts, the comprehension of which develops through four levels. Kafoussi (2004) used this model to guide the development of children's instructional activities, and then the analysis of their understanding of probability.

In our view, the substantial amount of cognitive knowledge on informal statistical reasoning has the potential to guide development of effective strategies for improving the statistical understanding of students and also researchers.

7. CONTRIBUTIONS OF THE PRESCRIPTIVE TO THE DESCRIPTIVE

We argued above that research on intuitive statistical reasoning (the descriptive facet) can guide instructional recommendations, and the design of teaching strategies, materials, and tools (the prescriptive facet). Evaluation of these prescriptive procedures gives information about their practical effectiveness, and also tests the underlying descriptive theory, enhancing or weakening its validity. Prescriptive research can thus contribute to the descriptive. For example, Sedlmeier's (1999) training program mentioned above demonstrated how the descriptive facet can influence the prescriptive. Sedlmeier used evaluation of his training programs (part of the prescriptive facet) to test the descriptive theories, thus demonstrating the interplay between the descriptive and prescriptive facets of statistical cognition. Statistical classrooms are the arena where the prescriptive is introduced, and where descriptive theories can be tested and be refined by evaluating the influence of different training programs.

8. INTEGRATION OF THE THREE FACETS IN STATISTICAL EDUCATION AND PRACTICE: TWO EXAMPLES

Statistics textbooks are based on the normative facet, but often incorporate also the author's descriptive and prescriptive ideas. Teachers and statistical consultants often use, in addition to the textbook and their statistical expertise, various pedagogical strategies to help students and researchers understand statistical concepts. They may consider the intuitive perceptions students and clients have at the start. They thus call on their own descriptive and prescriptive ideas in their efforts to assist the students and researchers

achieve good understanding. Do these ideas reflect their clients' conceptions and misconceptions? Research evidence can shed light on the validity of these professionals' mis/conceptions.

8.1. CORRELATION BETWEEN TWO DICHOTOMOUS VARIABLES

Consider, as a first example, descriptive research on a basic topic: correlation between two dichotomous variables. We will mention the normative model, describe research results from the descriptive facet and discuss possible implications for teaching, thus illustrating the mutual influence of the three facets of statistical cognition, and the research needed for EBP.

Adults' perception of the correlation between dichotomous variables had been examined in a number of studies (Beyth-Marom, 1982; Jenkins & Ward, 1965; Shaklee & Tucker, 1980; Smedslund, 1963; Ward & Jenkins, 1965). Usually, participants were given pairs of values, and asked to estimate the direction and/or strength of the relationship between the variables. Most research evidence found estimates were biased relative to the normative measure, which is based on the difference between two conditional probabilities. Participants' estimates were a function of the way data were presented (trial by trial, as a list of data pairs, or in a summary table); of the instructions given; and of whether asymmetric or symmetric variables were used (Beyth-Marom, 1982).

Task instructions were varied because the experimenters recognised that technical and lay usage of correlation, and other terms, may be very different. The kind of explanation participants were given was found to influence the estimates given, thus indicating how sensitive people are to language usage.

The asymmetric-symmetric distinction refers to whether the two values of the variable were different or similar. In the asymmetric case (e.g., pneumonia vs. no pneumonia; symptom present vs. symptom absent) one value has a lower status than the other, whereas in the symmetric case (e.g., gender: male, female), the values have a similar status. In the asymmetric case, the name of the variable is like the name of one of its two values (variable 'pneumonia', values 'pneumonia' and 'no pneumonia'). With symmetric variables, the name of the variable ('gender') differs from the name of its two values ('male', 'female'). Furthermore, in the asymmetric case, the two values may be described as 'occurrence', 'non-occurrence'. A 'non-occurrence' or 'negative' event has less impact on people's attention than a positive event (Nisbett & Ross, 1980). Participants' perception of correlation was much more biased for asymmetric variables, for which they tended to perceive only one or two cells of the full 2×2 table that is required for the normative assessment of correlation. With symmetric variables they tended to take account of all four cells—although not necessarily using the correct formula.

This research on naïve perceptions of correlation has a number of pedagogical implications:

1. When trying to explain a statistical concept, like correlation, teachers as well as students have to be aware of any different connotations a term may have in day to day language and in statistics.
2. The comprehension of correlation depends on a clear perception of the difference between variables and values.
3. It may be better first to use symmetric variables and then, after students understand that all four cells are relevant for the assessment of correlation, present examples involving asymmetric variables. Discussion of those might highlight the importance of the alternative values ('no symptom', 'no pneumonia') in estimating correlations and in other statistical tasks; and

4. Table format should be used at first, because students understand this format best. Then, later, students can work with other data presentation formats, perhaps by creating the 2×2 table themselves.

Pedagogical implications such as these illustrate how D can make a valuable contribution to P. These implications should shape training; the training should then be evaluated, thereby testing the validity of students' intuitive conceptions and the validity of the pedagogical implications. This demonstrates the pathway by which P may influence D.

It is unfortunate, although valuable, that the examination of issues in terms of our three facets identifies gaps in research knowledge that is essential for the effective adoption of EBP.

8.2. CONFIDENCE INTERVALS

For our second example, CIs, we focus on statistical practice—in particular, the formulation of advice for researchers. We have already introduced this example, but here we offer a quick review of the three facets of CI research. Normative research includes CI theory in mathematical statistics, and presentations intended to help researchers use CIs for data analysis (e.g., Altman, Machin, Bryant, & Gardner, 2000). The descriptive facet includes study of how researchers understand and interpret CIs. Prescriptive research includes study of how best to improve statistical practice: It provides evidence about what graphical design for CI figures and what wording used to interpret CIs most successfully communicate research results.

Now consider what research is available on CIs. Normative information is abundant, and journals continue to publish further theoretical results and applied techniques. In stark contrast, there is very little descriptive evidence about people's CI thinking. Cumming, Williams, and Fidler (2004) found that many researchers in psychology, behavioural neuroscience, and medicine hold the misconception that a 95% CI is also a 95% prediction interval for a replication mean, whereas a 95% CI has an average 83% chance of including the mean of a replication experiment. Belia, Fidler, Williams, and Cumming (2005) reported evidence of further misconceptions widely held by researchers about 95% CIs. These are some of the very few examples of descriptive research on CIs.

The next step is to suggest improved guidance for researchers and better graphical conventions for presenting CIs in figures and study whether these improvements are effective in overcoming those misconceptions: That, of course, is prescriptive research.

Cumming and Finch (2005) described a number of *rules of eye*, intended as simple guidelines for interpretation of 95% CIs shown in figures. Use of these rules would overcome misconceptions identified by Belia et al. (2005). Cumming (2007) presented figures and simulations to illustrate the relation between CIs and p-values; these were also intended to overcome some of the problems identified by Belia et al. The rules of eye, and illustrations of how CIs and p-values relate, are within the prescriptive facet, but need to be evaluated and found effective to become part of that facet's contribution to the EBP of statistics.

Our CI example identifies the importance of all three facets and their interactions, and emphasises the deficiencies of current descriptive and prescriptive knowledge. Considering statistical practice, there is some descriptive research identifying problems, but almost no prescriptive research showing what changes in practice, or what guidance to researchers, can be effective in overcoming those problems. It is especially important to expand descriptive and prescriptive knowledge about CIs because, as we mentioned earlier, persisting criticism of NHST is leading to recommendations that CIs be much more widely used in psychology and other disciplines (Cumming & Fidler, in press;

Fidler & Cumming, 2008). This highly desirable reform of statistical practice is being hampered by lack of evidence about effective ways to overcome CI misconceptions.

9. BARRIERS TO INTEGRATION

Despite the mutual influence of the three facets of statistical cognition and the obvious need for integration, their current state of fragmentation is a major obstacle to building a cohesive evidence base for statistical practice and education. Below are some quotations that illustrate the fragmentation; they will be familiar to many readers.

1. "… the almost universal reliance on merely refuting the null hypothesis … is basically unsound, poor scientific strategy and one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817).
2. "The believer in the law of small numbers … rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal 'explanation' for any discrepancy" (Tversky & Kahneman, 1982, p. 29).
3. "[Statistical] power is neglected by psychologists because, given their typically mistaken understanding of statistical significance, it is an unnecessary concept" (Oakes, 1986, p. 83).
4. "Activities specifically designed to help develop students' statistical reasoning should be carefully integrated into statistics courses" (Garfield, 2002).
5. "Since it appears that in judging randomness, subjects attend to the complexity of sequences, it might be possible to foster a more intuitive, yet mathematically sound, conception of randomness if it is introduced via the complexity interpretation" (Falk & Konold, 1994, p. 10).

The first quotation is an example of the vast literature that advocates reform of statistical inferences practices, questioning the normative justification of NHST. The next two describe people's intuitive perceptions (laypersons' as well as experts') of three statistical concepts: sampling, statistical power, and statistical significance. The fourth makes a prescriptive recommendation about the teaching of statistics. The final quotation integrates descriptive and prescriptive lines of research about randomness.

Normative, descriptive and prescriptive lines of research often study the same substantive content (e.g., CIs, correlation, randomness). However, the three lines of research have different goals, and are usually carried out by different scholars and published in different types of journals.

The Goals. Normative research aims to progress statistical theory, descriptive research aims to understand informal statistical reasoning, and prescriptive research aims to develop and evaluate improved strategies for teaching and practicing statistics.

The Scholars. Who are the people involved in this immense activity? Statisticians and mathematicians develop the science of statistics and so are most often responsible for the normative perspective. Cognitive psychologists contribute descriptive knowledge by studying how people reason statistically, interpret statistical concepts, make sense of statistical data; they describe people's correct or incorrect intuitions. Psychologists and educators in general, and teachers of statistics in particular, are often involved in studies aimed at improving statistical reasoning by suggesting new tools and methods of instruction (usually suggested by educators) or de-biasing techniques to overcome misconceptions (usually recommended by psychologists).

The Journals. Normative issues are mostly published in statistical journals, or in journals that focus on statistics and research methods in a particular discipline.

Statisticians publish in statistical journals (e.g., *Statistical Science*), while psychologists who are interested in normative issues often publish in the specialized psychological journals (e.g., *Psychological Methods*). Descriptive research on statistical reasoning often appears in psychology journals (e.g., *Journal of Behavioral Decision Making*, *Cognitive Psychology*), while prescriptive research is mostly seen in specialist statistics education journals (e.g., *The Journal of Statistics Education*, *Statistics Education Research Journal*, *Teaching Statistics*). These journals not only publish prescriptive research, but also to some extent have an agenda of integrating descriptive research, which is a laudable goal.

Regardless of the debate over when the modern era of statistical theory began, it is obvious that the normative tradition has a much longer history than either the descriptive or prescriptive traditions. Descriptive research dates back at least to the 1960s (e.g., Rosenthal & Gaito, 1963; Peterson & Beach, 1967). Prescriptive research on statistical practice and education is more recent, with formal societies and dedicated journals dating from around the mid 1980s.

We have demonstrated how often research in each of the three facets depends on the others and influences them. The organizational and sociological barriers between the three lines of research need to be removed, if EBP is to be achieved.

10. CONCLUSIONS

We have proposed the term statistical cognition for an integrative field that incorporates three lines of research. Interaction between the normative and prescriptive facets may seem relatively straightforward, so we focussed attention on the mutual contributions of the normative and descriptive facets, and the descriptive and prescriptive facets.

We discussed how normative models have been used as standards to which intuitive statistical reasoning (identified by descriptive research) is compared, with mismatches of varying extents emerging. The normative thus serves as a theoretical framework for describing how people *should* perform statistical tasks. We discussed NHST, CIs and Bayesian Hypothesis Testing as examples of how descriptive research on people's perception of statistical concepts can affect the normative status of models; this illustrates the contribution of descriptive research to the normative facet of statistical cognition.

Descriptive research on statistical reasoning aims to describe cognitive processes and misconceptions, and to detect developmental barriers to statistical reasoning. It can thus guide prescriptive investigations designed to identify the most efficient statistical training program. Conversely, prescriptive research on the effectiveness of the training program and teaching strategies, and the cognitive changes they elicit, provides empirical tests of descriptive models of people's statistical reasoning, thus enhancing or weakening their validity.

We used correlation as an example of how a statistical concept can be studied from all three perspectives of statistical cognition—its normative status, how people interpret it, and how it should be presented and explained—in order to improve statistics education and advising. We used CIs as an example of how research in all three facets might contribute to improving statistical practice. Finally, we identified barriers that we believe have hampered the interactions and synergies that are needed for EBP.

EBP is the key to better statistical practice and statistics education. It offers a number of advantages that should motivate its widespread adoption. Successful EBP

- ensures that consumers (in the fields we are discussing: researchers and students) get the best a discipline can offer;

- improves the efficiency of use of scarce resources, notably the time of teachers and other professionals;
- draws out practical implications from existing research for teachers and practitioners;
- guides the planning of future research; and
- encourages future research to generate practical implications.

A fear about EBP is that it might lead to a mechanistic, one-size-fits-all approach that marginalises the expertise and judgment of the teacher or statistician. However, recall the definition we quoted of medical EBP: "the integration of best research evidence with clinical expertise and patient values." We endorse this approach to EBP, which makes explicit the need for relevant expertise—of the teacher, statistician, or researcher—to ensure that lessons from the research evidence are applied appropriately, for maximum effect in a particular situation. Many factors influence decision making in statistical teaching and statistical consulting: statistical theory, ideology, values, clients, and personality factors. Even so, EBP can flourish.

In medicine, EBP has been primarily concerned with encouraging practitioners to make more use of research evidence that is already available. By contrast, in education greater emphasis has been placed on the absence of good quality research that can support EBP (Hargreaves, 1996). Davies (1999) emphasized that EBP in education should both draw on evidence from existing world-wide research and literature on education, and also encourage and guide further educational research.

In statistics education, there is already considerable descriptive and prescriptive research, and some integration of these two—for example in *SERJ*. However, there are also many gaps in the evidence needed to guide and justify EBP, and great scope for improved integration. It is not surprising that new research fields develop and specialize, building their own institutions, journals, and cultures. However, to adopt EBP in a thorough way requires reintegration, which both facilitates mutual contributions, and helps identify serious gaps in current knowledge. Reintegration is essential—and we believe the umbrella of *statistical cognition* can be very helpful—for building a cohesive and complete evidence base.

Why should labelling this integration and introducing the umbrella term help? It may, for example, help remind researchers, as they embark on a prescriptive or normative research program, to think also of relevant descriptive research that impacts on their goals—and vice versa. The term 'statistical cognition' in particular highlights the importance of a cognitive evidence base as well as a statistical and a pedagogical one.

For statistical practice, and especially for the reformed practice needed in psychology and other disciplines still over-reliant on NHST, the descriptive evidence base is very sparse, and very little prescriptive research has been conducted. There is enormous scope for a statistical cognition perspective to encourage and guide research, and to build the integrated evidence base needed for improved statistical practice and statistics education. The organizational and sociological factors responsible for the barriers between the three facets should now be exploited to overcome them. We look forward to further discussion of statistical cognition—and perhaps to the emergence of an international conference and a journal titled *Statistical Cognition*—and the potential we believe it has to energise and support the expansion of EBP in statistical practice and statistics education.

ACKNOWLEDGMENTS

This research was conducted while Ruth Beyth-Marom was a Visiting Fellow at the Institute of Advanced Studies, La Trobe University.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence* (2nd ed.). London: British Medical Journal.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923.

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance, 24*, 245-257.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology, 65*, 1119-1131.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389-396.

Bell, D. E., Raiffa, H., & Tversky, A. (1988). *Decision making: Descriptive, normative, and prescriptive interactions.* New York: Cambridge University Press.

Ben-Zvi, D. & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking.* Dordrecht, The Netherlands: Kluwer.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition, 10*, 511-519.

Beyth-Marom, R. (1990). Mis/understanding diagnosticity: Direction and magnitude of change. In K. Borcherding, O. L. Larichev & D. M. Mesick (Eds.), *Contemporary issues in decision making* (pp. 203-223). North Holland: Elsevier.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology, 45*, 1185-1195.

Biggs, J., & Collis, K. (1991). Multimodal learning and the quality of intelligent behavior. In H. Rowe (Ed.), *Intelligence, Reconceptualization and Measurement* (pp. 57-76). Hillsdale, NJ: Erlbaum.

Brown, J. S., Collings, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-41

Cockcroft, W. H. (1982). *Mathematics counts: Report of the committee of inquiry into the teaching of mathematics in schools*. London: HMSO.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*, 997-1003.

Cohen, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition, 7*, 385-407.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics, 29*, 89-93.

Cumming, G., & Fidler, F. (in press). The new stats: Effect sizes and confidence intervals. In G. R. Hancock & R. O. Mueller (Eds.) *Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers*. Hillsdale, NJ: Erlbaum.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170-180.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311.

Davies, H. T. O. (1999). What is evidence based education? *British Journal of Educational Studies, 47,* 108-121.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (Eds.). (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.). *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). New York: Cambridge University Press.

Edwards, A. L. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.). *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

Engel Clough, E., & Wood-Robinson, C. (1985). How secondary students interpret instances of biological adaptation. *Journal of Biology Education, 19*, 125-130.

Erickson, G. L. (1980). Children's viewpoints of heat: A second look. *Science Education, 64*, 323-336.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83-96.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory and Psychology, 5,* 75-98.

Falk, R., & Konold, C. (1994). Random means hard to digest. *Focus on Learning Problems in Mathematics, 16*, 2-12.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

Fidler, F., & Cumming, G. (2008). The new stats: Attitudes for the twenty-first century. In J. W. Osborne (Ed.). *Best practices in quantitative methods* (pp. 1-12)**.** Thousand Oaks, CA: Sage.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., et al. (2005). Toward improved statistical reporting in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology, 73*, 136-143.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239-260.

Fisher, K. (1985). A misconception in biology: Amino acids and translation. *Journal of Research in Science Teaching, 22*, 53-62.

Gal, I. (2002). Adults' statistical literacy: Meanings, components and responsibilities. *International Statistical Review, 70*, 1-25.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/garfield.html]

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal of Research in Mathematics Education, 19*, 44-63.

Garfield, J., & Gal, I. (1999), Teaching and assessing statistical reasoning. In L. Stiff (Ed.), *Developing mathematical reasoning in grades K*-12 (pp. 207-219). Reston, VA: National Council Teachers of Mathematics.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education 10*(2).
[Online: www.amstat.org/publications/jse/v10n2/garfield.html]

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond heuristics and biases. *European Review of Social Psychology, 2*, 83-115**.**

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgments.* Cambridge, UK: Cambridge University Press.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology, 40*, 631-666.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). San Diego, CA: Academic Press.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*, 1-20.

Hargreaves, C. (1996). *Teaching as a research based profession: Possibilities and prospects*. London: Teacher Training Agency.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118*, 248-271.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement, 60,* 661-681.

Huff, D. (1973). *How to lie with statistics*. Harmondsworth: Penguin.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 19*, 1-17.

Jones, G. A., Langrall, C. W, Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics, 32*, 101-125.

Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1999). Students' probabilistic thinking in instruction. *Journal for Research in Mathematics Education, 30*, 487-519.

Kafoussi, S. (2004). Can kindergarten children be successfully involved in probabilistic tasks? *Statistics Education Research Journal, 3*(1), 29-39. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(1)_kafoussi.pdf]

Kahneman, D. (2003). A perspective on intuitive judgment and choice: Maps of bounded rationality. *American Psychologist, 58*, 697-720.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). *Overcoming the inverse probability fallacy: A comparison of two teaching interventions*. Manuscript in preparation.

Kitson, A., Harvey, G., & McCormack, B. (1998). Enabling the implementation of evidence based practice: A conceptual framework. *Quality in Health Care, 17*,149-158.

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Lawrenz, F. (1986). Misconceptions of physical science concepts among elementary school teachers. *School Science and Mathematics, 86*, 654-660.

Leron, U., & Hazzan, O. (2006). The rationality debate: Application of cognitive psychology to mathematics education. *Educational Studies in Mathematics*, *62,* 105-126.

Meehl, P. (1954). *Clinical versus statistical prediction.* Minneapolis, MN: University of Minnesota Press.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.

Mehlinger, H. D. (1995). *School reform in the information age.* Bloomington, IN: Indiana University Press.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*, 23-63.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Oxford English Dictionary. (2007). Retrieved on October 15, 2007 from http://www.oed.com

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29-46.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. London: Routledge & Kegan Paul.

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. New York: Viking Penguin.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255-269.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175-197). Hillsdale, NJ: Lawrence Erlbaum.

Schmidt, F. L. & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-63). Hillsdale, NJ: Lawrence Erlbaum.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications.* London: LEA publishers.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 107*, 309-316.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgment of covariation between events. *Memory & Cognition, 8*, 459-467.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3-22.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165-173.

Smith, E. L., & Anderson, C. W. (1984). Plants as producers: A case study of elementary science teaching. *Journal of Research in Science Teaching, 21*, 685-698.

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education, 6*, 1-12.
[Online: http://www.amstat.org/publications/jse/v6n3/smith.html]

Stanovich, K.E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

The American Heritage Dictionary the English Language (4th ed.). (2000). Boston: Houghton Mufflin.

Trinder, L., & Reynolds, S. (Eds.) (2000). *Evidence-based practice: A critical appraisal.* London: Blackwell Science.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*, 1124-31.

Tversky, A., & Kahneman, D. (1982). Belief in the law of small numbers. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 23-31). New York: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*, 1-8.

Ward, W., & Jenkins, H. (1965). The display of information and judgment of contingency. *Canadian Journal of Psychology, 19*, 231-241.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46. [Online: www.stat.auckland.ac.nz/~iase/serj/SERJ2(2)_Watson_Callingham.pdf]

Webster's Online. (2007). Retrieved Oct. 15, 2007 from www.websters-dictionary-online.org

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

RUTH BEYTH-MAROM
Department of Education and Psychology
The Open University of Israel
108 Ravutski St. Raanana, 43107 Israel