Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2011)
174, *Part* 2, *pp.* 247–295

Towards more accessible conceptions of statistical inference

C. J. Wild, M. Pfannkuch and M. Regan

University of Auckland, New Zealand

and N. J. Horton

Smith College, Northampton, USA

[*Read before* The Royal Statistical Society *on Wednesday, October 20th, 2010, the President,* Professor D. J. Hand, *in the Chair*]

Summary. There is a compelling case, based on research in statistics education, for first courses in statistical inference to be underpinned by a staged development path. Preferably over a number of years, students should begin working with precursor forms of statistical inference, much earlier than they now do. A side benefit is giving younger students more straightforward and more satisfying ways of answering interesting real world questions. We discuss the issues that are involved in formulating precursor versions of inference and then present some specific and highly visual proposals. These build on novel ways of experiencing sampling variation and have intuitive connections to the standard formal methods of making inferences in first university courses in statistics. Our proposal uses visual comparisons to enable the inferential step to be made without taking the eyes off relevant graphs of the data. This allows the time and conceptual distances between questions, data and conclusions to be minimized, so that the most critical linkages can be made. Our approach was devised for use in high schools but is also relevant to adult education and some introductory tertiary courses.

Keywords: Computer animations; Informal inference; Sampling variation; Statistical inference; Statistics education

1. Introduction

This paper concerns the staged development of the big ideas of statistical inference over a period of years. It was prompted by the authors' need to make inferential ideas accessible to New Zealand school students aged approximately 14–17 years but much of its discussion is also relevant to adult education and introductory statistics courses at colleges and universities. The audiences that we most wish to engage include academic and professional statisticians. The other desired audiences for this paper are researchers in statistics education and teachers. Our biggest difficulty in trying to engage academic and professional statisticians on topics such as this is a common attitude that says, 'We don't care about school stuff'. We shall confront this before proceeding further.

We have often heard academic statisticians complain that statistics at school level is taught poorly, that most mathematics teachers have little or no training in statistics and that the statistics that is taught at school turns students off. 'What's the point? It all has to be redone from scratch at university anyway. Schools should just concentrate on laying solid mathematical

Address for correspondence: C. J. Wild, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand.
E-mail: c.wild@auckland.ac.nz

 0964–1998/11/174247

foundations that we can build on at university.' There are good reasons why academic and professional statisticians should care deeply about building more and better statistics into school curricula. Change is afoot and major curriculum developments are under way in many countries. We are not talking about continuing 'business as usual'. What statisticians should be heavily involved in is reconceiving what a more exciting and valuable school level statistics course could look like.

Why should statisticians care? First, there is the inherent but underexploited value of our product. There is a treasury of life skills lessons within statistics of value in the future lives of students regardless of what they end up doing. Second, there are dangers for society in statisticians failing to engage. Can any statistician really believe that it is desirable that society and its decision makers be made up of people whose minds have been conditioned by years of relentless determinism and who have no facility in stochastic thinking and no appreciation of its benefits? Third, there is the future of the discipline. Graduate programmes in statistics have traditionally relied on the conversion of people who started in mathematics undergraduate programmes. The declining numbers entering mathematics programmes in many countries mean that this strategy can no longer be relied on. We must build interest in statistics before students decide what to major in at university. In most jurisdictions that means that we must build interest in statistics while students are still at school. If we do not do this, then we must somehow grab the attention of someone who has no awareness of what statistics has to offer, and who is probably already planning to go into another area, and then reverse that decision in our favour. We have not shown any particular talent for such conversions in the past, so why would we bet the future of our discipline on such an implausible long shot?

Technology provides exciting possibilities for changing the landscape of statistics education in schools in ways that could make it unrecognizable. The inspirational 'Technology, entertainment, design' lectures of Hans Rosling (which are available from `http://www.ted.com/`), in which complex stories involving multi-dimensional data were made accessible to a general audience by using clever graphics, made this abundantly clear. In the same vein, work done in the SMART Centre of Durham University (Ridgway *et al.*, 2007a,b) shows that, with suitable visualization tools, ordinary teenagers can uncover and understand patterns involving interactions in four- or five-dimensional data. For statistics education, technology is the ultimate game changer. Its biggest pedagogical implications come from the fact that it allows us to conceptualize, in ways that were previously unavailable, potentially providing access to concepts at much earlier stages of development. With creative approaches, school level statistics can become much more ambitious, exciting and useful. Determining what a changed landscape could look like will, however, require the creative engagement of both academia and the profession. Although there are kernels of truth behind the objections some academic statisticians raise about school level statistics, these truths are simply evidence that there are difficulties which will require a large amount of creativity to overcome, that we need more and deeper engagement from more and better thinkers. The difficulties should not be taken as a justification for abandoning the battle field.

In a paper read before the Royal Statistical Society, Holmes (2003) brilliantly chronicled the history of statistics teaching in English schools and extracted important lessons to be learned from it. Holmes's main interest was in the journey towards statistics as a *practical subject* taught in *practical contexts* for *practical use*, to use some of his recurring phrases. This is statistics taught to enable students to understand better the real world that they live in, and sooner rather than later, in contrast with a rarefied enterprise that simply lays mathematical building blocks for future use. A prescient report of a Royal Statistical Society committee chaired by E. S. Pearson (Royal Statistical Society, 1952) was a major early milestone on this journey.

Unfortunately, it had to wait nearly 30 years for any serious implementation via the Schools Council Project on Statistical Education (1975–1980), which was led by Peter Holmes himself. The Schools Council Project in turn helped to inform the American Statistical Association's influential Quantitative Literacy Project of the 1980s, which was a watershed for parallel developments in the USA (R. L. Scheaffer, personal communication; see also Scheaffer (1990)). The report of the Statistics Focus Group sponsored by the Mathematical Association of America's Curriculum Action Project in 1991 was similarly a watershed for related developments for introductory courses at universities. The major thrust of the Focus Group's recommendations have survived to form the basis of the six recommendations that were fleshed out in the American Statistical Association's 2005 'Guidelines for assessment and instruction in statistics education' (GAISE) college report, namely: emphasize statistical literacy and develop statistical thinking; use real data; stress conceptual understanding rather than mere knowledge of procedures; foster active learning in the classroom; use technology for developing conceptual understanding and analysing data; and use assessments to improve and evaluate student learning. These sentiments also pervade the GAISE pre-K-12 report (Franklin *et al.*, 2007). Both reports are available from `http://www.amstat.org/education/gaise/`. We see our developments as next steps on this same journey.

Current realities fall far short of the worthy goals that were developed in the Royal Statistical Society and American Statistical Association. School level statistics for the bulk of students has suffered and stagnated for many years under a computational mentality pejoratively termed 'meanmedianmode' and the 'construct a graph' syndrome (Friel *et al.*, 2006). This has further been compounded by 'univariatitis' (Shaughnessy, 1997; Wild, 2006) and a focus on the construction of the tools of statistics rather than statistical reasoning processes resulting in a discipline that is perceived by many students and teachers as boring with little intellectual substance (Ridgway *et al.*, 2007a). Often descriptive statistics has been the only diet for students up to the penultimate year of high school, to be followed by an attempt to force-feed statistical inference, with its mathematical underpinnings, concepts and reasoning in the final year. It has not been entirely this way at all times in all places, but this has been the general tendency apart from 'an occasional creative oasis in a largely empty desert' (adapting Scheaffer (2002)).

The increased use of real data addressing interesting problems and multivariate data sets that permit students themselves to come up with interesting differences and other relationships to investigate are important new trends. The fostering of student engagement in data exploration as a 'data detective' (exploratory data analysis) is a hugely positive development in statistics education that partially obviates the problems above. Within it, however, lie the seeds of a new problem. When investigating interesting questions, relationships seen in data lead naturally to wanting to draw conclusions that apply to a universe beyond the data. Put more concisely, data addressing *motivationally compelling* questions beg inferences. Preventing inferential extrapolation makes the whole statistical exercise seem pointless. But, although good data and good questions make students want to make inferential claims, they currently have no rational bases on which to do so until they finally encounter formal inference. Moreover, the research on 'informal inference' that was reviewed in Wild *et al.* (2010) shows that students tend to grasp at it in incoherent ways. When the students do make claims they, and often their teachers, have no clear idea about whether they concern the data or a parent population. Additionally, research strongly suggests (e.g. Chance *et al.* (2004)) that large numbers of students fail to comprehend formal statistical inference when they do meet it at either school or introductory university level, and that they will continue to do so unless a much better job is done of laying essential conceptual foundations over a period of years before any attempt to teach formal inference is made. Otherwise there are simply too many ideas to be comprehended and interlinked all at once.

Work on informal inference has been going on in the statistics education research community, as a result of the statistical reasoning, thinking and literacy series of biennial international research forums that were initiated by Joan Garfield and Dani Ben-Zvi in 1999. Initially the forums addressed different types of statistical reasoning but the researchers at the 2005 forum came to a consensus that students should be learning to make inferences, initially informally. Consequently, the fifth forum in 2007 was focused on informal statistical inference (see, for example, the articles by Pratt and Ainley, Rossman, Pratt *et al.*, Zieffler *et al.*, Watson, Paparistodemou and Meletiou-Mavrotheris, Beyth-Marom *et al.* and Bakker *et al.* in volume 7, number 2, 2008, of the *Statistics Education Research Journal* (`http://www.stat.auckland.ac.nz/serj`) and by Makar and Rubin in volume 8, number 1, 2009). Konold and Kazak (2008) commented that the recognition that students need deeper understandings of inference is a move towards an acceptance that chance or sampling behaviour must be addressed. It is also noteworthy in view of the developments of this paper that the 2011 forum is to focus on new approaches to developing reasoning about samples and sampling in the context of informal statistical inference. For a review of this literature and some of its antecedents, see Wild *et al.* (2010).

Our own challenge, which has been made urgent by the demands of the imminent roll-out of the new statistics curriculum in New Zealand, has been to devise simpler versions of statistical inference in a way that lays solid conceptual foundations on which to build more formal inference in the longer term while giving students simple inferential tools, with reasonable operating properties, that they can use immediately. Moreover, we wanted the concepts to be built in a staged manner over several years so that there is time for them to be revisited several times to begin to bed in properly.

The remainder of this paper does not attempt to address broad aspects of statistics, of what should be taught and when, nor to illustrate exploratory data analysis with real data. It concerns educational experiences that *specifically target* statistical inference and, within that context, the development of integrated conceptual schema and tools that will assist students in making inferences when they are exploring real and interesting data.

2. Goals and pedagogical principles

2.1. Preliminaries

> 'Statistical inference moves beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are uncertain' (Moore (2007), page xxviii).

As is conventional in statistics, we employ the term 'statistical inference' to refer to the territory that is addressed by confidence intervals, critical values, p-values and posterior distributions. It addresses a particular type of uncertainty, namely that caused by having data from random samples rather than having complete knowledge of entire populations, processes or distributions. We shall add consideration of random assignment later. It does not address study design and execution issues, quality of data, relevance of data, practical importance and so on, even though we must pay attention to all these elements to employ statistical inference productively in the real world. This paper focuses simply on building the ideas of statistical inference in its conventional sense. And, since the focus is on building a particular set of concepts that have proved to be a real challenge in the past, we must proceed in a way that reduces competing distractions.

A number of discussions to follow will refer to Fig. 1, which displays data on the heights of a sample of 30 boys and of a sample of 30 girls aged 12 years taken from the CensusAtSchool

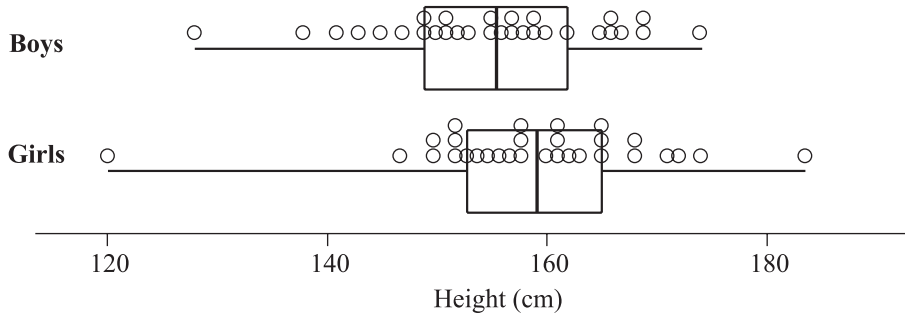


Fig. 1. Comparing the heights of boys and girls at age 12 years

New Zealand database. This data display combines a boxplot and dot plot for reasons that are given in Section 2.6. The motivation for comparing the heights of boys and girls at this age is the teacher folklore that says that, by maturing earlier, girls tend to be taller than boys for a short window of time at around this age. The median height for girls aged 12 years in our database is 1 cm higher than for boys. This is not detectable with even quite large samples. The apparent effect in samples of size 30 has boys taller than girls for almost 50% of samples taken. We can also achieve something very much like this by comparing heights of samples of 13- and 14-year-old girls, for example, where frequently the sample of younger girls will look taller on average. This can be useful in creating a so-called cognitive dissonance between what students see in their data and what they know to be true, and then exploiting that as a motivation for inference.

2.2. Goals

We want to arrive at conceptions of statistical inference that are accessible to the bulk of students and not merely an intellectual *élite*. We need to build on what is known from the research about students' intuitive reasoning. Since most students are not at all adept at abstraction we should try to keep everything we do as concrete as possible. By 'concrete' we mean closely anchored to actual physical experiences, or to simple ideas and representations that are already well understood by the vast majority of students. We need to identify *a minimal set of the biggest ideas of statistical inference* and to integrate inference for beginners within a holistic view of the investigative cycle (Wild and Pfannkuch (1999), section 2). Why do we need a minimal set of big ideas?: simply because learners cannot successfully address too many issues simultaneously. For beginners, we should therefore restrict ourselves to attempting to make just those few conceptual connections that matter most. To do this, we must reduce the mental clutter and time separation between the things that we most want to connect, avoid the distractions of competing issues that lead to cognitive overload and eliminate 'busy work' (see Wild (2007), section 2, for discussion and references).

Our basic approach consists of putting an approximate big picture securely in place initially and then, over a period of years, iteratively refining the details, adding subtleties, and even making corrections. Students are quite accustomed to this strategy as it is widely used in science for gradually building conceptions of complex realities. This dovetails with the strategy that was proposed in the GAISE report, which builds on the concepts of variability in levels A, B and C, as part of a developmental process that is based on statistical literacy, not age (Franklin *et al.* (2007), page 13). Level B is intended to help students with 'interpreting results with an

eye toward inference to a population' (page 58), whereas level C begins to describe how to draw conclusions from data.

Although the motivation and building of inferential ideas take substantial time, we want this development to lead to an implementation in which the equivalent of a statistical significance decision can be made extremely quickly. We believe that, in the context of a particular investigation a student is doing, *the mechanics of the inferential step should not be at all demanding*. Whenever students have to struggle with details of implementation the big picture becomes lost (Cobb (1997), page 80), taking with it any consciousness of why they are doing what they are doing and what it all means once they have finished. Ideally, supplementing slower detailed experiences such as project work, we want students to be able to have many experiences where they can ask a question, get the data, obtain the graphs, make descriptive comments and inferential conjectures, make a call on something like the direction of a group difference and write up the whole story—all within a one-hour lesson period. Students need repeated experiences in which all these big steps can be gone through in a very short period of time for the big picture to be seen as a whole and for the connections between its main elements to be made strongly. Such an approach also emphasizes that the inference step is not 'what statistics is all about', but one small step in a much larger, deeper and richer set of statistical activities focused on making sense of the world.

Our ideal is for a wholly *visual approach* backed up (see Pfannkuch *et al.* (2010)) by language that reinforces and communicates the essence of what is being seen, experienced and thought about. Our visual approach will attempt to minimize the conceptual distances between the concrete realities, including precursor practical experiences, and the dynamic imagery envisaged. Our ideal, moreover, has the inferential step *able to be performed without students taking their eyes off their graphs* so that the connections between question, data and answers are kept as immediate and obvious as possible. At worst we want any 'eyes off the graph' statistical processing to be reduced to an absolute minimum. We believe that any inferential guidelines proposed should have reasonable operating properties in repeated sampling. And, because our aim is *to develop a pathway* to formal inference, it follows that the conceptions that are arrived at *should have intuitive connections to the more formal methods to be used later*.

In summary:

(a) we should work from a minimal set of the biggest ideas of statistical inference;
(b) the *mechanics* of the inferential step should not be at all demanding;
(c) inferences should be able to be performed without students taking their eyes off their graphs;
(d) methods should have connections to the more formal methods to be used later.

We assume a stage of development in which the basic ideas of centre have already been established and students have already experienced dot plots and boxplots.

2.3. Description versus inference

We cannot build statistical inference without first building an appreciation of sample *versus* population, of description *versus* inference, and of characteristics of samples giving us estimates of characteristics of populations. Regarding sample *versus* population, the more generally useful conception of the distribution that is generated by a process, rather than population, is probably too subtle for beginners (Wild, 2006). Adapting the analogy of Pratt *et al.* (2008) of two games being played in statistics, we also need to start from a clear distinction between when we are playing the *description game* and when we are playing the *inference game*. Almost anyone with any statistical training, when looking at plots like Fig. 1, will ask themselves 'Are these two

things different?'. In doing so they are jumping straight into inference. Back in the description game the answer to the question is plainly obvious. Of course they are different and different in a myriad of ways, even if they might sometimes look very similar overall. 'What game am I currently playing?' matters greatly because different games have different goals and different rules.

2.4. Sampling variation: the raison d'être of statistical inference

Although these elements provide some of the necessary background for inference, the critical element—the element that statistical inference was developed to confront—is sampling variation. Thus, any conceptual approach to statistical inference must flow from some essential understandings about the nature and behaviour of sampling variation.

The English language connotation of 'sampling variation' is the idea that we see something different every time that we take a new sample. In other words, it suggests sample-to-sample differences rather than the fact that the samples misrepresent their parent populations to a greater or lesser extent. When we make inferences we are trying to allow for the uncertainty due to having data from a sample rather than having the whole population (or process, or distribution). But we obtain our picture of the extent to which sample data tend to represent and misrepresent a parent population by looking at the properties of characteristics of the data (e.g. means) over the repeated taking of samples, i.e. by investigating the patterns of sampling variation. The phrase 'uncertainty due to sampling variation' is actually a code for uncertainty due to the fact that we have sampled, where the degree of uncertainty we should allow for is estimated by using lessons learned from studying patterns of sampling variation.

Research has shown (see Chance *et al.* (2004) and Makar and Confrey (2004)) that the experiences that we have been giving of sampling variation of means, including computer animations, are very difficult to grasp. It is virtually impossible to make students reliably summon them up when looking at something like Fig. 1. This has led some to suggest that sampling variation is too difficult to teach to beginners. However, in the context of a classroom where every student obtains their own personal sample from a given population, it is obvious to everyone involved that everyone will obtain different graphs. So it is not the idea of sampling variation *per se* that is difficult to grasp. The problems that have so far proved intractable are the problems of building within students a reliable propensity to conjure up the ideas of sampling variation whenever they look at something like Fig. 1.

2.5. What belongs in other sets of experiences and what can we leave until later?

What can we strip away so that the biggest issues that are addressed by statistical inference are left exposed? Since statistical inference (used in its conventional sense) is designed to deal with uncertainties about the true state of nature due to sampling variation, we believe that experiences that are designed to build and cement the ideas of statistical inference *should focus solely on sampling variation*. Other types of variation (see section 3 of Wild and Pfannkuch (1999)) such as 'random' measurement error, although statistically indistinguishable in some important models, are unwarranted complications. Issues of 'Am I measuring the right thing?' are very different yet again and, together with non-sampling errors and issues of relevance and quality of data, belong in complementary learning modules in which the main focus is on planning and critiquing investigations and not on introducing the ideas of statistical inference. The issues of experiment *versus* observational study, or causation *versus* association, though crucially important, are sufficiently tangential to the concerns of formal statistical inference that they should be targeted separately. Issues of inference in exploratory *versus* confirmatory settings

are directly related but are embellishments for much later in a student's development. Checking of assumptions is important for formal inference but, since no distributional assumptions are explicitly made in our precursor forms of inference, checking assumptions also falls into the future embellishments category. Any distinction between practical and statistical significance should also be postponed. Trying to do it all at once creates confusion and, logically, this distinction can only be drawn for someone who already knows something about what statistical significance is.

In reported research on informal inference, teaching experiences involving reasoning about differences in centres have often drawn on *context matter knowledge* (Watson, 2008) throwing yet another complication into what students are doing with description and inference. How we make formal statistical inferences draws only on patterns in data. Critiquing the plausibility of an inference draws on external knowledge of context as does any consideration of the practical importance of differences seen. But we should not throw all these things into the mix together too early. Our proposals for informal inference will draw solely on patterns in data.

Let us return to Fig. 1 and the 'Are they different?' question. It is the question that everyone who has taken a tertiary level statistics course learns to ask. When prompted, however, every statistician recognizes that it is generally not a meaningful question. In terms of populations it makes no sense at all. Of course they will be different if measured sufficiently accurately. Why would we *ever* expect the means of different populations to be exactly the same? Of course one will be bigger than the other. The questions that make sense are 'Which one is bigger?' and 'How much bigger?'. To be fair, 'are they different' makes rather more sense in a randomized experiment where it can be at least plausible that an experimental intervention makes no difference at all to the outcome under study—though who would ever do the study without a strong suspicion of a real difference?

So why is everyone taught to narrow in on this 'are they different?' question and to hypothesize 'there is no difference'? Largely, it has been a device that has let us calculate (or estimate) probabilities and to produce a numeric uncertainty measure. Unfortunately, 'suppose that there is no difference, calculate and then interpret something like a p-value' is not a mode of thinking that comes naturally to people. On the contrary, it is rather like looking at the world while standing on your head (for a discussion of student's difficulties, see sections 5–7 of Rossman (2008) and Cobb and Moore (1997), section 3.5). These considerations have led us to try to eliminate '*thinking under the null*' from our *beginning experiences of inference*. Where we cannot make a call on which is bigger because we are unsure of the direction of the population patterns, we say 'It is too close to call: *I can't tell* which is bigger'. In contrast, use of the terms difference or same encourages students, and often their teachers, to make misconceived claims that two populations are the same (accepting the null).

2.6. On concrete foundations

The proposals that we make in Section 3 are presented in terms of sampling from populations. They depend on boxplots providing a bridge between reasoning entirely from graphics to reasoning from summaries in ways that converge, qualitatively, to the two-sample t-test. They are motivated by using particular conceptions of sampling variation conveyed using new forms of computer animation. Everywhere along this journey there are dangers that students will become lost in mystifying abstractions. So, at all stages, we need to maximize linkages to concrete realities and then to things that can be easily seen in graphics. Our focus in this paper is on conceptual flow, not the details of classroom implementation. Our group has also been working on classroom implementations that emphasize discovery learning but this will be reported elsewhere (the first such paper is Arnold and Pfannkuch (2010)).

We *start with sampling finite populations*, such as the students in the databases of CensusAtSchool New Zealand (`http://www.censusatschool.org.nz/`). This is much more concrete for beginners than having to imagine a conceptual population or data being generated by a process. It is immeasurably more concrete than sampling from theoretical distributions such as the normal distribution. This does not stop us from applying the methods that are obtained for dealing with uncertainty to a broader range of data than finite population data, but the flow of inferential ideas is developed in this simple context.

Many misconceptions and errors in interpreting summary statistics arise because summary statistics are introduced in terms of algorithms and presented divorced from their role as summary features of distributions. We can keep reminding students of the data and distributions that the summary statistics summarize by continually presenting them as annotations of simple dot plots of data.

The forms of summary that relate in the most visually obvious way to the points that are depicted in a dot plot are the median, quartiles and extremes—in other words, the ingredients of the basic boxplot. It is very easy to guesstimate and draw a boxplot by hand on top of a dot plot and, indeed, doing so is probably the best way of gaining an appreciation of just what a boxplot actually is. The *box plot provides a natural bridge* between operating entirely in terms of what is seen in *graphics* to reasoning using *summaries*. In data analytic settings, we always present *boxplots* for beginners in conjunction with the *underlying dot plots* because the boxplot in isolation is a very abstract entity. Retaining the dots, as done in Fig. 1, provides a reminder that the boxplot is just summarizing the raw data, thus preserving a connection to more concrete foundations.

The developments that are proposed in Section 3 are built from particular conceptions of sampling variation built by using animated, computer-simulation-based graphics. This is the only feasible way to demonstrate the effects of sampling variation with very large numbers of repetitions. Used alone, simulation and the resulting animations can just be computer magic—not only unreal and unconvincing, but often not even understood (see Wild (2007)). Chance and Rossman (2006) emphasized the importance of starting with practical physical simulations which then become automated by using computer simulations as a means of ensuring that students understand fully what a computer animation is doing. Our classroom implementation work incorporates this strategy (Arnold and Pfannkuch, 2010).

3. A proposal for early forms of inference

3.1. A motivational metaphor

Our scene setting metaphor for statistical inference starts with the idea that looking at the world by using data is like looking through a window with ripples in the glass (Fig. 2). 'What I see in my data is not quite the way it really is back in the populations that they come from.' This fundamental idea must be internalized before statistical inference can make sense. The patterns to be seen in data are distorted versions of the patterns that are present in the populations or processes that they come from. On occasion, the distortions may even be so big that the patterns that we think we see are just artefacts caused by the ripples in the glass. To make inferences from data, we need an appreciation of how these distortions arise, when they are likely to be large and when they are likely to be small. Where beginners are engaged in a module that is focused on statistical inference, we limit attention to distortions that are produced by the act of sampling and sampling variation (Fig. 3).

Using physical experiences with sampling and sampling variation which lead into computer animation experiences, we seek to build the appreciation (Fig. 4) that with small samples we

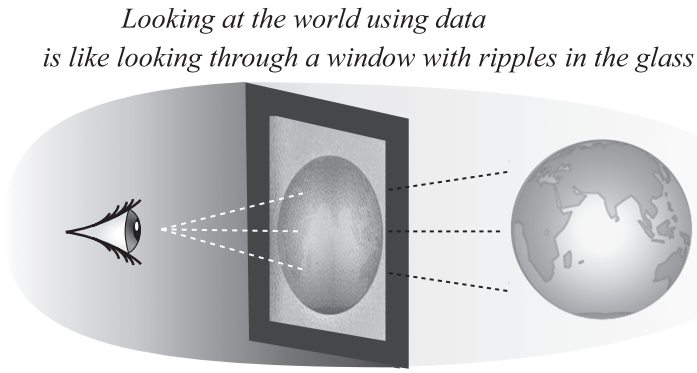


Fig. 2. 'What I see is not quite the way it really is'

Fig. 3. Distortions due to sampling

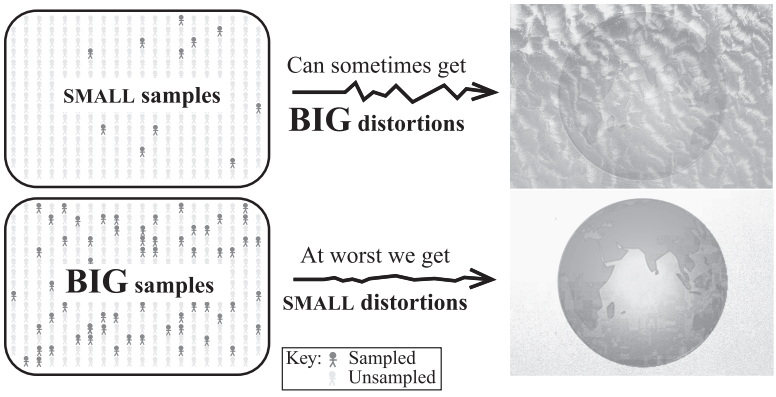


Fig. 4. Distortions related to sample size

can obtain big distortions (cf. very rippled glass) but with big samples we obtain only small distortions (cf. slightly rippled glass). Figs 2–4 are general metaphors for statistical inference.

3.2. A new visual approach to sampling variation

The next step is to link plots like Fig. 1 to depictions of sampling variation that occupy the same visual space as the target plots by using computer animations. Movement provides a powerful means of displaying the nature of variation. It also captures attention, which is a fact that the producers of on-line advertisements are increasingly exploiting. Because the experience of an animation is impossible to convey in a static, print, journal paper, a Web page has been set

up at `http://www.censusatschool.org.nz/2009/informal-inference/WPRH/` containing all the animations that are described in this paper. We shall also do our best to convey the main ideas verbally.

Simply animating the repeated sampling-and-display process, by briefly displaying each pair of samples sequentially, is an excellent first step (see panels 1(a), 1(b) and 2(a) of the Web page). Such animations show clearly the variation in centres, spreads, etc., as we take new samples, and the effect of sample size on these features (see panels 1(b) and 2(b) of the Web page). However, the frames of such an animation do not retain any memory of what has gone before, and so leave no lasting impression of the extent of the variability. In our follow-up animations, all the box plots that are seen over time leave behind 'footprints' with the most recent plot superimposed over the set of footprints. We use colour to distinguish between the current and historical boxes, and the median *versus* the rest of the box.

What builds up over time is images that look like those in Fig. 5 (animated in panel 3(b) of the Web page). In Fig. 5, since colour is not available to us, the boxplots from the past are grey with the medians printed somewhat darker than the rest of the box. What builds up is a blurred image with the latest box overprinted, printed in black. When animated, the black box appears to vibrate in position (and width) and leaves behind a record of the extent of the sampling variation in the medians and in the rest of the box. Students must be led to realize that, when in an investigation, they collect their own data and construct their own set of boxplots, what they have is the equivalent of a single frame from this movie. We note that although the effect of sample size is visible in Fig. 5 it is much more dramatically conveyed by Fig. 6 (see also panels 1(b) and 2(b) of the Web page). We urge the reader to go to the Web page because the effects of colour and movement are key ingredients of these displays.

To remind the students that the samples are being taken from populations while keeping the main emphasis on what is happening to the (animated) samples, we represent the populations in the top half of Fig. 5. At the beginning of the animation, just before the samples start to appear, we grey the populations out to push them into the visual background and label them 'the unseen world'. Once what is happening in the animations is understood, calling to mind sampling variation with boxplots reduces to recalling the vibrating boxplots as in Fig. 7, together with 'I have to take into account this sort of uncertainty about where the true boxes lie when I make

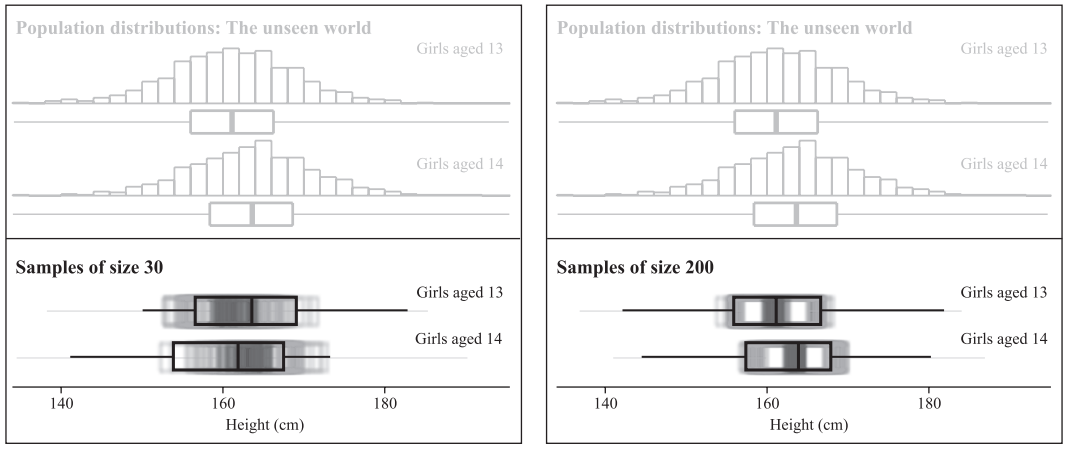


Fig. 5. Boxplots with a memory over repeated sampling (animations at `http://www.censusatschool.org.nz/2009/informal-inference/WPRH/`, panel 3(b))

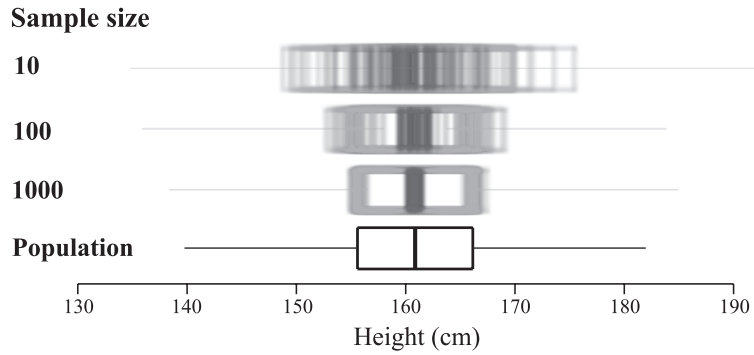


Fig. 6. Effect of sample size: sampling from a single population (animations at `http://www.censusatschool.org.nz/2009/informal-inference/WPRH/`, panel 2(b))

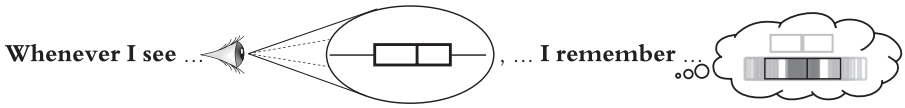


Fig. 7. Desired habit of mind (animation at `http://www.censusatschool.org.nz/2009/informal-inference/WPRH/`, paragraph 1)

my comparisons'. This is a considerably smaller conceptual leap than connecting the boxplots in Fig. 1 with traditional representations of the sampling error of the mean. The traditional representations occur in an entirely separate visual dimension from the data plots and thus have no obvious linkage to the plots that students are trying to interpret.

The animations show the effects of sample size clearly (see panels 2(a) and 2(b) of the Web page). Additionally, when we sample girls aged 13 and 14 years as in Fig. 5, everyone knows the direction of the true difference. As we step through the movies slowly frame by frame, we see that the sample direction of the difference is often opposite to the true direction when we have samples of moderate size, but virtually never in the wrong direction when we have large samples. The effect reversals in moderate-sized samples can be used to create a cognitive dissonance between what students see in the plots of the data and what they know to be true about the heights of 13-year-olds and 14-year-olds.

3.3. Making the call and estimating effect sizes

Can we conclude from Fig. 1 that girls tend to be taller than boys back in the populations that we have sampled from? We know that the boxes that we see are not quite in the right places ('what I see isn't quite the way it really is'). The images of sampling variation, especially the vibrating boxplots (Figs 5 and 6), experiences of samples telling the opposite story from what is actually happening back in the populations, or the opposite story to the plot of a neighbouring student, and experiences of the effect of sample size, lead us to the basic ideas that are depicted in Fig. 8. This diagram deals generically with 'Can I make the call that B's values tend to be larger than A's values back in the population(s)?'.

The basic idea underlying Fig. 8 is that we should only make the call if the location shift that we see between our boxes is sufficiently big to override the uncertainties that are illustrated in Fig. 7 about where the real boxes lie. The levels of uncertainty will be large with smaller samples and small with very large samples. If we do not feel that we can make the call, our answer is 'I do not have enough data to be able to tell which tends to be bigger', i.e. 'I can't tell'.

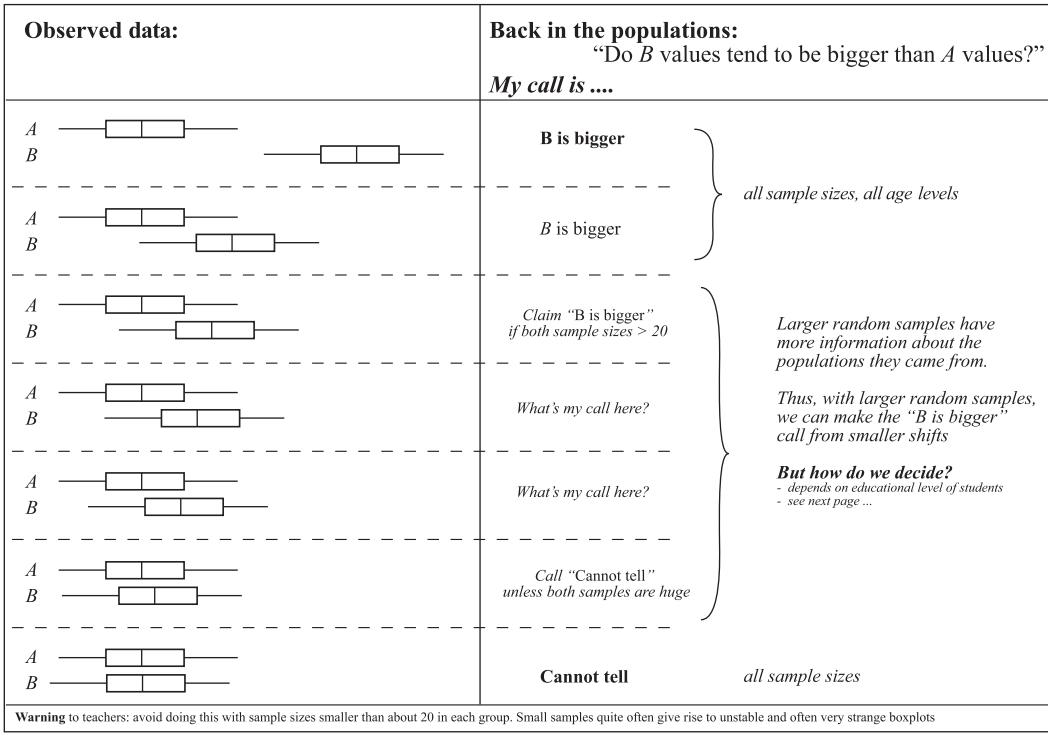


Fig. 8. When can I make the call that B tends to give larger values than A?

How big does the shift have to be (moving up Fig. 8) before it is sufficiently big for us to make the call? We need ways to operationalize this basic idea that are sufficiently simple for students to handle, involve the big ideas about the effects of spread and sample size, have reasonable operating properties in repeated sampling and can be refined over time to become increasingly more like methods that are accepted by statisticians.

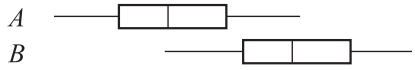
Although this paper simply proposes guidelines, our implementation work with Pip Arnold concentrates on leading students to discover for themselves that they need some sort of decision guideline and to come up with ideas for guidelines that approach those that are presented here. The decision guidelines that we have devised to support the new New Zealand high school statistics curriculum are depicted in Fig. 9 where one milestone will usually be targeted per year of schooling with milestone 4 occurring in the last year.

Since the guidelines will look unfamiliar, we give the reader a secure reference point by noting that the milestone 3 test is a very minor tweaking of Tukey's notched boxplots technique for making visual inferences (McGill *et al.*, 1978). When operating these guidelines for data analysis, students should be working from plots that are obtained from software. Because it is time consuming, drawing plots by hand is busy work that obstructs the mental connections that we are trying to make (and these are extremely straightforward to create by using appropriate technology).

Note that the conceptions that are depicted in Fig. 8 remain constant over all four levels and are continually reinforced. What changes in our proposal as we progress through the milestones (Fig. 9) is a gradual refinement of how to determine whether an observed shift is sufficiently large to make the call. Teachers should not feel constrained to follow through all the development levels that are depicted in Fig. 9. We stress that milestone levels can, and should, be skipped over

Guidelines on "how to make the call" by development level

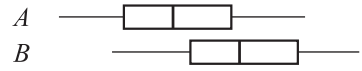
At all levels:



If there is no overlap of the boxes, or only a very small overlap
make the call immediately that ***B tends to be bigger than A*** back in the populations

Apply the following when the boxes do overlap ...

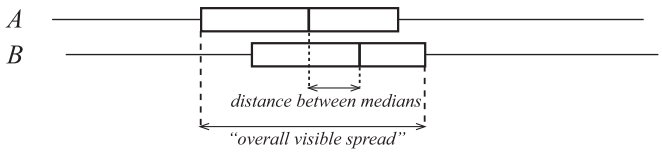
Milestone 1 test: *the 3/4-1/2 rule*



If the median for one of the samples lies outside the box for the other sample
(e.g. "*more than half of the B group are above three quarters of the A group*")
make the call that ***B tends to be bigger than A*** back in the populations

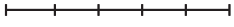
[Restrict to samples sizes of between 20 and 40 in each group]

Milestone 2 test: *distance between medians as proportion of "overall visible spread"*



Make the call that ***B tends to be bigger than A*** back in the populations
if the distance between medians is greater than about ...

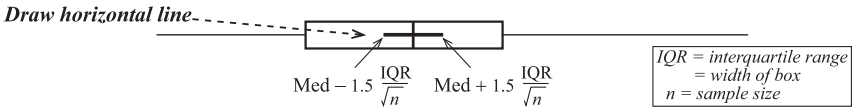
1/3 of overall visible spread for sample sizes of around **30**



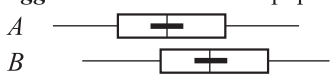
1/5 of overall visible spread for sample sizes of around **100**

[***Could also use*** 1/10 of overall visible spread for sample sizes of around 1000]

Milestone 3 test: *based on informal confidence intervals for the population median*



Make the call that ***B tends to be bigger than A*** back in the populations



if there is compete separation between the added intervals (i.e. do not overlap)

Milestone 4: ***on to formal inference***

Fig. 9. How to make the call by level of development

whenever the students already have the requisite statistical maturity. Here, we simply elaborate on the definitions of the guidelines with discussions of their rationale and operating properties being postponed until Section 3.4.

An intuitive take on the milestone 1 test is that we can make the call if the median for one sample lies beyond 'the great whack' of the other sample. It can be operated almost instantly from the graph. Sample size is not taken into account and teachers are asked to limit themselves to sample sizes of around 20–40. This has the advantage of simplifying the procedure at the cost of limiting its utility. This trade-off seems appropriate given the multiple concepts that are being introduced and developed.

To operate the milestone 2 test we advocate that students use the version that is closest to the sample sizes they have and simply do a quick freehand subdivision of a line representing total visible spread into thirds or fifths and make the call on that basis. We want the focus to be on the big ideas and do not want this to degenerate into an exercise about the accuracy of application of the $\frac{1}{3}$ and $\frac{1}{5}$ cut-offs.

It is a fairly short step from the vibrating boxplot versions of Figs 5 and 6 to putting some sort of interval of uncertainty around the data median to try to capture the population median. The formulae for milestone 3 are simple and students should calculate and add these lines three or four times to graphs that are produced by software. We would be entirely happy if this was done by using approximate values for the median, and box widths read off the graphs as the only reason for doing any hand calculation at all is to help to establish the idea. Subsequently, it is desirable that the annotations are put on by software. The intuitions to be appealed to for milestone 3 follow from thinking of these thick lines as uncertainty intervals—'I'm thinking that the true median is likely to be in here somewhere'. If there is no overlap between where I think true median B is and where I think the true median A is, then I can make the call.

The guidelines in this three-step sequence are sufficiently simple for students to operate. Milestones 1 and 2 can be operated with no 'eyes off the graph' processing at all, as can milestone 3 with modest assistance from software. It does not require much 'eyes off the graph' processing even in the absence of such assistance. The guidelines for milestone 2 and beyond do involve the big ideas about the effects of sample size and within-sample spread, the latter being finessed painlessly by the graphics.

Merely being able to make a call on the direction of a difference is a minimal and unsatisfying form of inference, however. The milestone 3 graphic and the intuition about 'where I think true medians A and B are' readily lend themselves to visually constructing an informal confidence interval for the true difference in population medians by using the method that is depicted in Fig. 10. (If there was overlap between the thick lines in Fig. 10, then the lower confidence limit would be negative.) This provides an intuitive underpinning for formal confidence intervals for differences in means at milestone 4, whether from procedures based on Student's *t* or resampling.

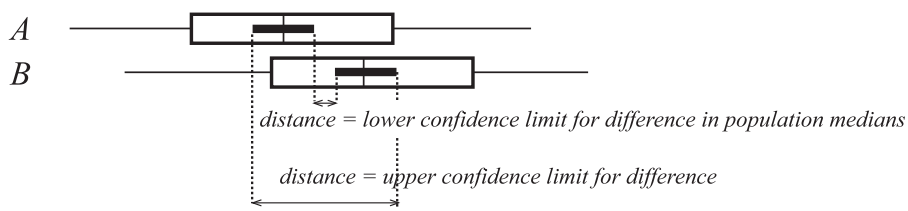


Fig. 10. Reading off a confidence interval for a true difference

Analysis tools implementing the graphics depicted in this paper are freely available from the CensusAtSchool New Zealand site (`http://www.censusatschool.org.nz/`) and the Web page `http://www.censusatschool.org.nz/2009/informal-inference/WPRH/`. The boxplots used for data analysis retain the underlying dot plots and have milestone 3 annotations as an option. We only omit the dots from animated boxplots and figures that are intended to convey ideas about how to read inferential information from the boxplots.

3.4. Rationales and justifications

3.4.1. Milestone 1

The big lessons about statistical inference that we want our students to take away at the milestone 1 level are

(a) that samples can give us useful approximate pictures of what is happening in populations,
(b) an ability to see approximate location shifts in dot plots or boxplots,
(c) an appreciation that the story that the data suggest about the population can be wrong (e.g. a sampling-variation-induced reversal of the reality),
(d) an appreciation that the shift that is seen in data must be reasonably substantial before we can fairly safely infer the direction of a population effect from the direction of a data effect and
(e) a simple way to implement the previous idea.

When statisticians look at something like Fig. 1, they take in the horizontal sweep and see the box, or set of points belonging to a single group, as an entity. Features like location shifts leap out at us immediately. There is research evidence (Cliff Konold, personal communication) that middle school students scan these plots quite differently from experts: that they want to look vertically and to compare details in one group with details in another. (The horizontal–vertical considerations are simply reversed if we run the plots the other way.) Research (e.g. Bakker *et al.* (2005)) has shown that students will fairly naturally compare the medians and quartiles, etc. of boxplots for one group with those of the other, and that it is not just corresponding features that they compare. We have leveraged off this for our milestone 1 level. For these first exposures we most want to begin to ease students into seeing shifts and to start bedding in the idea that we can make the call if the shift is sufficiently large relative to spread. This is aided by emphasizing overlap first, rather than differences in centres as the latter invites looking at differences absolutely and not relatively. We postpone making a feature of the effect of sample size until milestone 2, after the ability to see shift and overlap has already been well established. The milestone 1 level can be skipped over for groups of students who already have this ability.

Simulations with normal data give type I error rates for our milestone 1 test of approximately 15% for samples of size 20 in each group, 7% for samples of 30, 3% for samples of 40 and 0.4% for samples of size 100. So, under the guidelines, the students will be making the call roughly in line with conventional practices of statistical inference.

3.4.2. Unpacking the two-sample t-test

The basic idea of the two-sample t-test, or the Welch test, is to base a decision on significance on the distance between centres of the two samples expressed as a multiple of the standard error of this difference. The standard error of the difference is a combined measure of the spreads of the two samples deflated by square-root sample size. Equivalently, it makes the call if the distance

between centres as a proportion of within-sample spread exceeds a cut-off which depends on the sample sizes—with bigger cut-off values being used for smaller samples. Our decision guidelines now begin to converge towards this idea.

3.4.3. Milestone 2

At the milestone 2 level all of the milestone 1 points (a)–(e) should be reinforced. Two new ingredients are stressed at milestone 2: first, that sample size matters when making the call and, second, a moving of attention towards distance between centres as a proportion of a spread. Our first attempt at this guideline compared the distance between medians with the sum of the interquartile ranges but we were told by the teachers whom we were consulting that this was too difficult for their students and this conversation led us to the 'overall visible spread' idea that is shown in the diagram. We obtained the very simple cut-off proportions that are depicted by using simulations with normal data. The type I error rates are about 8% at the anchor sample sizes. There is a trade off between more conventional type I error rates at memorable sample sizes (30 is 'traditional classroom size') and having an extremely simple rule. We gave more weight to the latter. The round number sample sizes with approximate 5% type I error rates are $n = 40$ for $\frac{1}{3}$, $n = 80$ for $\frac{1}{4}$ and $n = 125$ for $\frac{1}{5}$. The type I error rates with data from the strongly skewed χ^2_4-distribution and the heavy tailed t_4-distribution are very similar to those from the normal distribution at the anchor sample sizes. Despite the milestone 2 guidelines being transitional as far as formal significance testing is concerned, they have lasting value as rough rules of thumb for exploratory data analysis.

3.4.4. Milestone 3

Milestone 3, which continues our convergence towards the big idea of the t-statistic, is a very minor modification of Tukey's notched boxplots idea. We have used a slightly smaller but eminently more memorable multiplier, namely 1.5. This increases the large sample type I error rate with normal data slightly from about 2% at moderate sample sizes to about 2.5% (with essentially identical behaviour for normal, χ^2_4- and t_4-distributions). Additionally, we have used a thick horizontal line in place of a notch. We think that this works better visually as a means of conveying uncertainty about where the true median lies. In fact the thick horizontal lines are approximate 90% confidence intervals.

Some statisticians may be uncomfortable about using the non-overlap of individual uncertainty intervals to indicate significance. After all, the reading of significance from the overlap *versus* non-overlap of confidence intervals is something that students and researchers who are outside statistics naturally want to do and something that many teachers of elementary courses at universities expend significant energy on stamping out. It is not that this procedure is actually wrong. It is simply that, by operating this way, you are working with type I error rates for significance tests for a difference that are much smaller than the coverage error rates for the individual parameters, a fact that is exemplified by the figures given in the previous paragraph. If you want these error rates to agree, you must do something slightly different. Regardless of what we do, teachers at more advanced levels will have to confront this issue anyway, whether in terms of error rates or the standard error of the difference not being the sum of the standard errors. We believe that this is a finer distinction, later refinement issue rather than a fundamental issue. And, when teachers do confront the issue, it will just be this one small issue that they can attend to on its own rather than having it mixed in with other bigger picture considerations.

3.4.5. Milestone 4

At milestone 4, which is beyond the scope of the present paper, we can bring in such things as the notion of a null hypothesis, levels of variational behaviour under the null due to sampling or randomization, normal distribution models, a change in emphasis for measures of location and spread from the median and interquartile range to the mean and standard deviation motivated by efficiency arguments under normal models, and formal methods of inference based on *t*-tests, randomization or resampling.

3.4.6. Historical parallels

Chris Triggs has drawn our attention to parallels between our work here and a literature in the 1940s and 1950s on simpler forms of inference, and in particular Tukey (1959) in which he presented a simpler alternative to the two-sample *t*-test. This test was based on the number of observations in sample A that are below all observations in sample B plus the number in sample B that are above all in sample A and associated statistical tables were provided. The following quotations are particularly interesting in the current context,

> '... the needs of certain users for such a procedure which would be very much easier to use (and teach) than those so far available.'

> 'A "pocket test" of the present sort has rather definite uses. It is for use "as a footrule", "on the floor", "in the field" etc.'

> 'Simplicity means practical portability—the ability of the statistician to carry the procedure everywhere, stored in a very small part of his memory.'

3.5. Extensions

3.5.1. Categorical variables

When students have become accustomed to the milestone 3 method for making the call in Fig. 9 and can read off confidence intervals for effect sizes as in Fig. 10, we can then extend the same mode of thinking to bar charts for categorical variables. The graphs in Fig. 11 relate to repeated sampling from the database and using the 'travel-to-school' variable. The animation is produced as follows. Whenever we take a new sample we marked the positions of each of the tops of the most recent set of bars with a horizontal blue line (grey in Fig. 11). Over time this leaves the set of 'footprints' that are shown in Fig. 11. When played fast the black box tops appear to vibrate (vertically) over the accumulating set of footprints. Fig. 11 is from the final frame of our 'animation movies' in which the boxes from the population percentages are superimposed on top of the footprints.

This suggests superimposing uncertainty intervals when we obtain a set of bars from a single set of data obtained in a study as in Fig. 12(a). These are drawn (by using methods to be discussed later) to enable students to make a call on 'which is bigger' when comparing categories and to obtain approximate confidence intervals for differences; cf. Fig. 10. We can do the same sort of thing when comparing two samples as in Fig. 12(b). This lets us compare, for example, the proportions of Auckland students who ride bicycles to school with the proportion of Christchurch students who ride bicycles. The latter is obviously considerably larger, even allowing for sampling variation. One possible explanatory factor is that Auckland is a hilly city and Christchurch has a much flatter terrain.

The intervals of uncertainty that are drawn on Figs 12(a) and 12(b) are not the usual standard error bars or confidence intervals that are often added to such plots. Their lengths are calculated so that visual confidence intervals for differences that are obtained as in Fig. 10

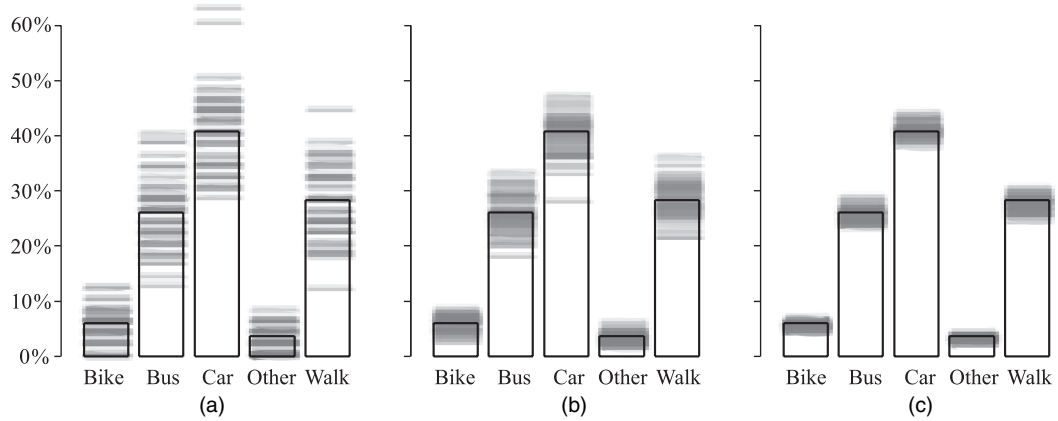


Fig. 11. Sampling variation with a single-category variable, 'How do you travel to school?' (animations at `http://www.censusatschool.org.nz/2009/informal-inference/WPRH/`, panel 4(a)): (a) samples of size 50; (b) samples of size 200; (c) samples of size 1000

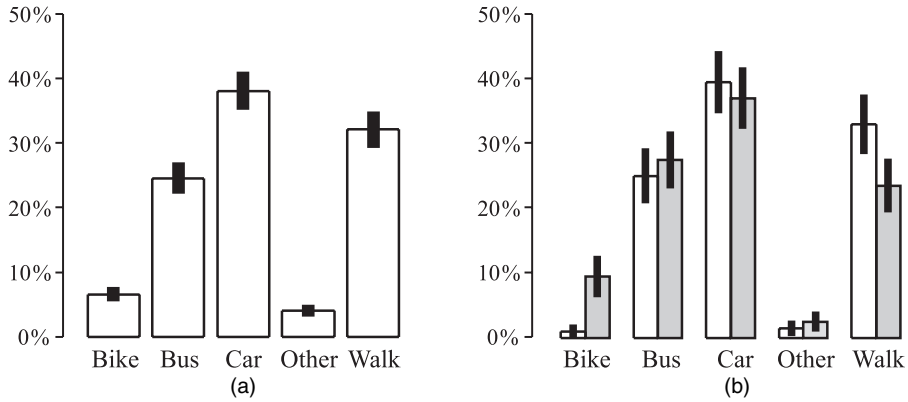


Fig. 12. Bar charts with intervals of uncertainty designed to address differences: (a) comparisons within a sample; (b) comparing two samples (□, Auckland; ■, Christchurch)

approximate the usual normal-approximation-based intervals by using a method to be published elsewhere. Since there is no intuition that is accessible to high school students in our construction of the intervals, they should just be added by software. Nevertheless, a crude version attaches $\pm 1.5\sqrt{\{p(1-p)/n\}}$ to boxes as in Fig. 12(b). Here, the natural comparisons are things like comparing the proportions of bikers between Auckland and Christchurch, namely proportions from separate or independent samples. The $\pm 1.5\sqrt{\{p(1-p)/n\}}$-formula could be presented for brighter students with attention drawn to the way that the intervals grow smaller as p approaches 0 or 1 and, as the sample size increases, mimicking the behaviour that we see in Fig. 11. Obtaining intervals for Fig. 12(a) is much less straightforward. Inferences for this setting are seldom presented even in introductory statistics at university level because these proportions are not independent, so we have the complications of accounting for the correlation structure. In fact, the crude attachment of $\pm 2.2[\sqrt{\{p(1-p)/n\}} - 0.1/\sqrt{n}]$ works surprisingly well for almost all comparisons except those involving very small probabilities (for which the standard asymptotic intervals perform badly anyway). There are on-line analysis tools that implement these methods at `http://www.censusatschool.org.nz/`, in the iNZight

data analysis system (`http://www.stat.auckland.ac.nz/~wild/iNZight`), and an R library available from `http://www.stat.auckland.ac.nz/~wild/VisDiffs`.

3.5.2. Randomized experiments

Repeated random assignments can produce very similar patterns of variation to random sampling when viewed with graphics like Fig. 5. We are perfectly comfortable with students at this stage of their development applying the methods that were developed for random sampling to experimental data. This puts us in the same company as all those who apply t-tests and F-tests to experimental data: a rather large company! Our intention is to postpone any focus on contrasting random sampling with random assignment until milestone 4. For a diagram concisely relating the features of random sampling, random assignment and the consequent scope of inference, see Ramsey and Shafer (2002), page 9.

4. Discussion

Many of the problems with students learning statistics stem from too many concepts having to be operationalized almost simultaneously. Mastering and interlinking many concepts is too much for most students to cope with. We cannot throw out large numbers of messages in a short timeframe, no matter how critical they are to good statistical practice, and hope to achieve anything but confusion. We need complexity reduction strategies. One of these is to cluster concepts into smaller, more manageable sets that share fairly well-defined spheres of influence. In this paper we have concentrated on the cluster of concepts relating to statistical inference (in the traditional sense of the term). Even here our main worry is about too many ideas rather than too few. Of course, a cluster of concepts cannot stand on its own, no matter how well it is linked internally. We also need strategies to build bridges between clusters or to activate the right cluster at the right time, but that is beyond the scope of this paper.

In current approaches to informal inference, the arguments that are used as 'evidence' by students are often incoherent and are not part of a planned incremental development. Informal inference should not just be a matter of doing whatever you can towards reaching an ambitious goal without scaffolding. Turning students loose can be an excellent approach for research into the thinking patterns of students at various stages of their development, but it is not pedagogy and does not help them to learn to marshal more coherent forms of evidence. Any approach to inference that is accessible to beginners will inevitably be 'too simple', however. Techniques will often be pushed beyond the limits of where they work well when students try to obtain satisfying answers to real and interesting problems—as they must to make statistics vibrant! Additionally, no accessible approach can address all the issues that a professional would address. But all of this is acceptable, we believe, so long as we are building important intuitions and are on a planned development path to something better.

A possible criticism of the approach that is taken here is that 'students should not be taught material they then have to unlearn'. Our sequence of guidelines for making a call may seem to violate this. However, our guidelines do not have to be unlearned. Each is statistically valid but with a limited range of applicability. We move on from one level to the next to extend the range of applicability. The guidelines will be useful for the rest of their lives. For example, the milestone 2 guideline gives useful visual cues when looking at plots that are not supplemented by inferential information. Every professional's statistical life is a voyage of discovery, of 'that's fine as far as it goes but it has its limitations', running into those limitations and then finding a way forward. We believe that instructional experiences should mirror this. If rules for living

statistically are simply received 'carved onto tablets of stone', any seemingly capricious smashing and replacement of those tablets is disturbing. But that does not apply to lessons that are extracted as part of a discovery learning process. Ingredients for such a process can include the sequence need → idea → does it work? → simulate and see → if it seems to work then use it, followed up later by encountering a set of situations where a methodology is clearly no longer working as it should, leading to ideas about how to proceed, sending us right back along the previous sequence of steps.

When we have started to explain our work, some friends and colleagues asked 'Why not just do randomization tests and bootstrap resampling?' (cf. Cobb (2007) and Rossman (2008)). 'Surely that solves everything. They involve very little machinery.' We demur from this as a recipe for the earliest experiences for the following reasons. Conducting randomization tests involves an appreciable amount of time with the eyes off the data. It involves going off into another thinking paradigm and returning again. But, most importantly, these tests have the intrinsically difficult ideas of 'thinking under the null' at their conceptual heart. This is further compounded when we construct confidence intervals, and there is much less discussion by randomization proponents of this essential element. Obtaining confidence intervals is a process that requires marshalling the conceptual mysteries of inverting a test. This makes randomization inference intrinsically more difficult than our proposals so our preference is for randomization inference to follow the developments here, and that is the way that it has been staged in New Zealand's new curriculum. We have similar reservations about introducing the bootstrap too early. It seems to us to be a recipe for confusion to stir ideas of sampling from the sample into the mix at the very point when students are only just getting to grips with the ideas and implications of sampling from a population. Randomization and bootstrap inference will be introduced at the final year of high school (milestone 4). We are currently working towards this.

We set out to devise a developmental path beginning early in high school that would lay some solid intuitive and big picture conceptual foundations to be built on in their final high school year when they are taught formal methods of making statistical inferences. In terms of the practical problems that can be addressed by using our methods at milestone 3, we have managed to go almost as far as the average first undergraduate statistics course by using only a very small number of fairly simple ideas. We have been able to avoid 'thinking under the null'. Our highly visual approach to 'making the call' and obtaining confidence intervals can be operated so quickly that they should at most be a minor impediment to experiencing all the major steps of the investigative cycle, including writing about what has been learned, in a short period of time. How well the guidelines work is something for students to investigate by using simulation.

The presentation in Section 3 concentrated on visual precursor forms of statistical inference and the nature of the guidelines, which are basically decision rules. Many statistics education researchers dislike rules intensely and want to distance statistics education from them. When it comes down to making a 'statistical significance' type of call on whether B tends to give bigger values than A, however, or to provide some sort of interval estimate, the methods that real statisticians use are essentially rules, be they Bayesian or frequentist, based on parametric assumptions or randomizations. The disaffection of statistics education researchers with rules comes, we believe, from a proper horror at widespread teaching practices that are aimed at 'getting students through the test' via a path of least resistance. These lead to the blind application of rules, which are uninformed by any insight about what is being done and why, and with all vestiges of common sense disengaged. In other words, the root worries are about rules as a replacement for thinking, even as a barrier to thinking, rather than rules as aids to thinking that can help to marshal our common sense. In New Zealand we are mounting a three-pronged attack on this root problem. First, we are devising discovery learning pathways leading to a realization

of a need for guidelines, ideas about what they might look like and including investigation of operating properties. Second, we are embedding practical operation of the rules as one small part within a holistic approach to communicating about data (see Pfannkuch *et al.* (2010)). Third, new national assessments being devised will make it impossible to obtain good marks by blindly applying tests. High marks will require demonstrating understanding and insight.

We began this paper with an appeal for more academic and professional statisticians to draw inspiration from gifted communicators of data stories like Hans Rosling and to involve themselves in reconceiving what a more fascinating, valuable and ambitious school level statistics could look like. Computer technology has changed the world entirely. So let us try our hands at working from essentially blank slates and see whether we can come up with creatively new ways in which students can interact with and learn from data, and new ways of conceptualizing the big ideas of statistics. We have attempted a little of that here and hope that many others will join in this enterprise.

Acknowledgements

The authors are grateful for helpful comments on drafts of this paper from Alan Agresti, Alasdair Noble, Anthony Harradine, Arthur Bakker, Bill Finzer, Cliff Konold, Richard Scheaffer, Ilze Ziedins, Joan Garfield, Mike Camden, Rob Gould, Roxy Peck, Sandy Madden, Sandy Pollatsek, Tom Louis, the editorial team and the referees. These acknowledgements should not, however, be taken to imply that all those listed are in complete agreement with what we have written. This work was partially supported by a grant from New Zealand's 'Teaching and learning research initiative' (`http://www.tlri.org.nz`).

References

Arnold, P. and Pfannkuch, M. (2010) Enhancing students' inferential reasoning: from hands on to "movie snapshots". In *Proc. 8th Int. Conf. Teaching Statistics* (ed. C. Reading). The Hague: International Statistical Institute. (Available from `http://www.stat.auckland.ac.nz/~iase/publications.php`.)

Bakker, A., Biehler, R. and Konold, C. (2005) Should young students learn about boxplots? In *Proc. International Association for Statistical Education Roundtable on Curricular Development in Statistics Education* (eds G. Burrill and M. Camden), pp. 163–173. Voorburg: International Statistical Institute. (Available from `http://www.stat.auckland.ac.nz/~iase/publications.php`.)

Chance, B., delMas, R. and Garfield, J. (2004) Reasoning about sampling distributions. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (eds D. Ben-Zvi and J. Garfield), pp. 295–324. Dordrecht: Kluwer.

Chance, B. and Rossman, A. (2006) Using simulation to teach and learn statistics. In *Proc. 7th Int. Conf. Teaching Statistics* (eds A. Rossman and B. Chance). Voorburg: International Statistical Institute. (Available from `http://www.stat.auckland.ac.nz/~iase/publications.php`.)

Cobb, G. W. (1997) Mere literacy is not enough. In *Why Numbers Count: Quantitative Literacy for Tomorrow's America* (ed. L. A. Steyn), pp. 75–90. New York: College Entrance Examination Board.

Cobb, G. W. (2007) The introductory statistics course: a Ptolemaic curriculum? *Technol. Innovns Statist. Educ.*, **1**, 1–15. (Available from `http://escholarship.org/uc/uclastat_cts_tise`.)

Cobb, G. W. and Moore, D. S. (1997) Mathematics, statistics and teaching. *Am. Math. Mnthly*, **104**, 801–823.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M. and Scheaffer, R. (2007) *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: a Pre-k-12 Curriculum Framework*. Alexandria: American Statistical Association.

Friel, S., O'Connor, W. and Mamer, J. (2006) More than "Meanmedianmode" and a bar graph: what's needed to have a statistical conversation? In *Thinking and Reasoning with Data and Chance: Sixty-eighth Yearbook* (eds G. Burill and P. Elliott), pp. 117–137. Reston: National Council of Teachers of Mathematics.

Holmes, P. (2003) 50 years of statistics teaching in English schools: some milestones (with discussion). *Statistician*, **52**, 439–474.

Konold, C. and Kazak, S. (2008) Reconnecting data and chance. *Technol. Innovns Statist. Educ.*, **2**, no. 1. (Available from `http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1/`.)

Makar, K. and Confrey, J. (2004) Secondary teachers' statistical reasoning in comparing two groups. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (eds D. Ben-Zvi and J. Garfield), pp. 147–168. Dordrecht: Kluwer.
McGill, R., Tukey, J. and Larsen, W. (1978) Variations of box plots. *Am. Statistn*, **32**, 12–16.
Moore, D. (2007) *The Basic Practice of Statistics*, 4th edn. New York: Freeman.
Pfannkuch, M., Regan, M., Wild, C. J. and Horton, N. (2010) Telling data stories: essential dialogues for comparative reasoning. *J. Statist. Educ.*, **18**, no. 1. (Available from `http://www.amstat.org/publications/jse/v18n1/pfannkuch.pdf`.)
Pratt, D., Johnston-Wilder, P., Ainley, J. and Mason, J. (2008) Local and global thinking in statistical inference. *Statist. Educ. Res. J.*, **7**, 107–129.
Ramsey, F. L. and Schafer, D. W. (2002) *The Statistical Sleuth*. Pacific Grove: Duxbury.
Ridgway, J., Nicholson, J. and McCusker, S. (2007a) Teaching statistics—despite its applications. *Teachng Statist.*, **29**, no. 2, 44–48.
Ridgway, J., Nicholson, J. and McCusker, S. (2007b) Reasoning with multivariate evidence. *Int. Elect. J. Math. Educ.*, **2**, 245–269.
Rossman, A. (2008) Reasoning about informal statistical inference: a statistician's view. *Statist. Educ. Res. J.*, **7**, no. 2, 5–19.
Royal Statistical Society (1952) The teaching of statistics in schools. *J. R. Statist. Soc.* A, **115**, 126–137.
Scheaffer, R. L. (1990) The ASA-NCTM Quantitative Literacy Project: an overview. In *Proc. 3rd Int. Conf. Teaching Statistics* (ed. D. Vere-Jones). Voorburg: International Statistical Institute. (Available from `http://www.stat.auckland.ac.nz/~iase/publications.php`.)
Scheaffer, R. L. (2002) Statistical bridges. *J. Am. Statist. Ass.*, **97**, 1–7.
Shaughnessy, J. M. (1997) Missed opportunities in research on the teaching and learning of data and chance. In *Proc. 20th A. Conf. Mathematics Education Research Group of Australasia* (eds F. Biddulph and K. Carr), vol. 1, pp. 6–22. Sydney: Mathematics Education Research Group of Australasia.
Tukey, J. W. (1959) A quick, compact, two-sample test to Duckworth's specification. *Technometrics*, **1**, 31–48.
Watson, J. M. (2008) Exploring beginning inference with novice grade 7 students. *Statist. Educ. Res. J.*, **7**, no. 2, 59–82.
Wild, C. J. (2006) The concept of distribution. *Statist. Educ. Res. J.*, **5**, no. 2, 10–26.
Wild, C. J. (2007) Virtual environments and the acceleration of experiential learning. *Int. Statist. Rev.*, **75**, 322–335.
Wild, C. J. and Pfannkuch, M. (1999) Statistical thinking in empirical enquiry (with discussion). *Int. Statist. Rev.*, **67**, 223–265.
Wild, C. J., Pfannkuch, M., Regan, M. and Horton, N. J. (2010) Inferential reasoning: learning to "make a call" in theory. In *Proc. 8th Int. Conf. Teaching Statistics* (ed. C. Reading). The Hague: International Statistical Institute. (Available from `http://www.stat.auckland.ac.nz/~iase/publications.php`.)

Discussion on the paper by Wild, Pfannkuch, Regan and Horton

John MacInnes (*University of Edinburgh*)

This paper is important for three reasons. First it asks us to think about statistics education in an original but fundamental way: what is the simplest, clearest, but theoretically sound way to teach inference—a keystone idea of the discipline—to students who may have little or no facility with mathematics, nor any particular interest in statistics nor any desire or ambition to study it further? Second it provides an opportunity for us to think about the relevance of their approach in teaching statistics to other audiences. Finally, it ought to make us think also about school mathematics education and its fitness for preparing students, at school, college or university, to understand statistics.

I have complete sympathy for the argument of Wild and his colleagues that statistical ideas, including inference, can usefully be taught in schools, and that schools should not be seen as providing only core mathematics competencies. I am a sociologist, but I find that colleagues in the life and even the natural sciences often find that the mathematics competencies that are taught in UK schools focus too much on passing assessments in mathematics techniques and far too little on either their application or any sense of the relevance and power of numbers.

I share also their desire to put the relationship between samples and populations at the heart of the matter; their faith in 'complexity reduction strategies' (which is not a bad definition for the whole enterprise of statistics) and their insistence on 'motivationally compelling learning' rather than proceeding logically and abstractly from first analytic principles. Univariate descriptive statistics might constitute, logically, the ground floor of the statistical edifice, but students often find it such an uninspiring place that they seldom want to explore the rest of the building. I teach postgraduate students who have little or no previous statistical training. They also struggle to get a clear idea of a sample and population, simply because they are

used to proceeding on the assumption that data always come from that mythical beast the 'representative' sample. I thus endorse the authors' decision to expel 'external context knowledge' from the learning process.

I share their realism about students' inability to 'think under the null', their enthusiasm for visual approaches and above all their belief that

> 'rules too often serve as replacement for thinking, even as a barrier to thinking, rather than ... aids to thinking that can help to marshal our common sense'

(school mathematics again?). The students I teach also struggle to 'think under the null' and, presented with precisely the variety and complexity of messages that the authors describe so well, can come to grasp at formulaic rules as a lifebelt, rather than developing an understanding through exploring data and coming to realize both the potential and the limits of a sample.

I have three, small, reservations. The first concerns their wish to keep such a clear distinction between 'the description game' and 'the inference game'. There are costs and benefits here. Sufficiently accurate descriptions might inevitably yield differences, but I find that a clear distinction between description and inference is one that beginners neither see nor need to focus on: they describe in order to infer. Their aim is usually inference from a sample to a population, and description of their data is the means to achieve it. What is fundamental, as the authors argue, is 'building within students a reliable propensity to conjure up the ideas of sampling variation' when they see data. Comparing shift and spread is a good way to get at the mechanics of inference, but both come, in a sense, from the 'description game'.

Second, is there a danger that students might draw a correct but misleading conclusion from the milestones: that big samples are good, full stop, rather than bigger samples make finer comparisons possible? Will they not see our aim, to use the authors' visual metaphor, to see the world through the clearest possible glass?

Third, and I doubt that the authors would disagree, having secured some knowledge of inference, I would want to pay at least some attention to other sources of error, especially measurement error. At least in the social sciences, the construction of data is something that I find students struggle with. They find it easy to write off almost any attempt at measurement as hopelessly 'relativist'. However, once data are in their hands, they rapidly acquire an aura of almost Papal infallibility. Measurement is another key source of 'ripples' in the glass.

Wild and his colleagues argue that 'technology is the ultimate game changer' in statistics education, making a convincing case for the role of computer-generated animations in teaching. Innovative visualization of data is an indispensable weapon, not only in teaching statistics but also in securing greater general statistical literacy. However, good animations need rare computing skills and time to produce. They do not fit well either with the standard artisan model of university education, nor with the existing system of funding it. This may be something that can be done better, and ought to be funded, collectively.

Finally, let me make an observation about what we might call the generation of statistical knowledge and awareness. Let us distinguish

(a) the development of the discipline of statistics in its own right,
(b) its application by those formally trained in statistics and
(c) its use and application by others.

This threefold classification is imprecise because the boundary between groups (b) and (c) is porous. The development of statistics has always been tied to its application, from agriculture to astronomy. However, with the expansion, professionalization and bureaucratization (in its Weberian, rather than any pejorative, sense) of scientific enquiry, many of those using and applying statistics will have studied the discipline, if at all, briefly and in a purely applied context. It is a tool that they may use with less concern for its inner workings than whether 'it does what it says on the tin'. It is also imprecise because group (c) is virtually limitless. As the Royal Statistical Society's 'getstats' campaign reminds us, any active, and even not so active, citizen can benefit from statistics in making sense of an ever increasing supply of public data, or assessing risk.

We too often assume that the relationship between these three groups is simple and hierarchical. Professional statisticians whose expertise is rooted in their immersion in the discipline should set standards for, and where possible deliver, the training of the second group, as well as taking some responsibility for enlightening the third. Thus the Royal Statistical Society's mission is

(a) to 'nurture the discipline of statistics by publishing a Journal, organising meetings, setting and maintaining professional standards, accrediting university courses and operating examinations' and

(b) to 'promote the discipline of statistics by disseminating and encouraging statistical knowledge and good practice with both producers and consumers of statistics, and in society at large'.

In making such an assumption we perhaps pay insufficient attention to two issues. First there is no necessary association between having specialist knowledge and being able to describe it. This applies with particular force when its description necessarily involves a novel specialist language for the learner. Second, the relationship between our three groups works best when hierarchy is moderated by vigorous communication. Perhaps the most important lesson of this paper is that we can best advance statistical knowledge when we listen carefully to those who do not yet have it.

Peter Holmes (*Sheffield*)

I am pleased to second the vote of thanks to our authors for a very interesting paper. In particular there are three innovative insights and use of software that I commend. These are

(a) the insight of the sample as seeing the population through a distorting lens,
(b) the software that retains the different results of a boxplot with fading so that you have the idea that when I see one I remember the variability pattern and
(c) the idea of making a call that x is larger than y which hides the formal requirement for a null hypothesis.

So why do I feel distinctly uneasy when reading the paper as a whole?

It took me back to when I first taught school statistics in 1959 to students who were 16 years old and older. Here it was assumed that all data were good data and that all samples could be considered as random samples. These assumptions were not valid then; even less are they valid now.

The emphasis in the paper is on laying the foundations of classical statistical inference based on random samples. This is only one aspect of what is being done in modern day statistics. The authors say that these things are for all learners—just developing specific techniques over time. But all this takes time and all school time is under pressure. You would have to justify this particular set of teaching over other things within the school curriculum. Even when you have made the case for teaching statistics to all you have to justify which statistics teaching to which pupils. You have to prioritize. The authors dismiss other aspects of statistics—yet many of them are more important for most students.

The only distorting effect of the lens (the sample) that is considered is that of variability. A far larger distorting effect in most of the data that the students are likely to come across or to generate themselves is that of bias. All students need to be aware of how to obtain good data and to judge whether the data they have are good data.

The idea of making a call comes down to a set of rules. The learner could go away with the impression that the more statistical techniques you know the closer the two samples can be for you to make a call. Yet this is *not* what we want. You can always make a call if the two samples are different (e.g. in means or medians)—the important thing is to have some idea about how likely you are to be wrong. The idea that you might be wrong is fundamental and is seriously understated in this paper.

For most learners it is more important to know that, in the circumstances where you have random samples, when you make a call the following make you more likely to be right:

(a) the greater the difference between the samples;
(b) the larger the samples (so smaller between-sample variation);
(c) the smaller the variability within the sample.

Only then move on to the principle that we could use things like medians and edges of boxplots to indicate an acceptable degree of certainty in making this call.

This emphasis on random samples and sample variability raises the question of what is important in school statistics. The work of statisticians is to obtain and interpret data. Many applications today have data sets that do not come easily under the category of representative samples from a population. Many are population statistics in their own right; some could be samples—but if they are so seen it is not clear what sort of sample they are (e.g. sales by a supermarket to its customers using loyalty cards). It is the job of the statistician to make sense of these data. It is an important part of school statistics that students become aware of the sort of problems that statisticians work on and the sort of inferences they can draw.

This brings us to the idea of statistical inference. The authors take statistical inference to mean only the sorts of inference drawn in classical statistics—drawing inferences from random samples or those that can be considered as random. Statistical inference is much, much more than this. Intuitive statistical inference is broader than laying the basic foundations for classical statistical inference.

It starts when you look at the data and ask *what do these data tell me?* In early school years this can lead to straight deductive answers (e.g. more pupils in our class come to school by bus than by car). But it easily moves to inference when the question is changed to *what might have caused these data to be as they are?* This is the start of inference.

Answers to this question might include references to variability as in this paper, but they also give rise to many other possibilities and important statistical ideas. This question raises the idea of accuracy of the data—how were they measured and obtained? So we consider the nature of questionnaires and questionnaire design or collecting data efficiently through experimental design. It also raises the nature of the sample (if it is a sample) from which the data were drawn. This question may also point us to looking at similar data from the past to see how things have changed and inferences about what might have caused them to change. These raise the whole idea of bias in the data as a possible explanation for what is there. They are also topics that are readily grasped by most secondary school students.

All this links more with scientific understanding of inference and possible sociological or scientific explanations—and these might be open to further investigation on which is more or less likely. In my view these are more important aspects of statistics that all students should know and, if there is insufficient time to teach all, should take precedence over the more rule-oriented decision making that is described in the later parts of this paper.

The vote of thanks was passed by acclamation.

Adrian Bowman (*University of Glasgow*)

I add my thanks to the authors for a very stimulating paper. It must be recognized that there has sometimes been a slightly uneasy relationship between the world that is inhabited by teachers, of any subject and at any level, and the world of educational research. The authors of this paper are to be commended for helping to bridge the gap for statistics. In particular, they have clearly demonstrated the benefit of understanding what goes on in the minds of those who meet statistical concepts for the first time.

I agree with the authors that the development of concepts in the same 'visual space' is very helpful. This includes not only concepts of data but of models, e.g. in two-way analysis of variance where the superimposition of fitted means from different models on carefully structured plots of the data can help greatly in explaining ideas, in particular of interaction.

I agree with the authors that animation is a hugely valuable tool. It is not yet widely recognized that animation is actually readily available in R which has become to a large extent a common 'computational space' for statisticians. One example is the `rpanel` package (Bowman *et al*., 2007), which aims to make the construction of animations as easy as possible to do, particularly in an everyday, time-pressured teaching context.

These principles are applicable in more sophisticated contexts. For example, in non-parametric regression animation allows us to view the effects of altering the complexity of the model that we fit. There are standard methods of indicating uncertainty in our estimation, but graphical reference bands, expressing where our estimate is expected to lie if the relationship is linear, can provide an effective assessment of the suitability of that simple model. The authors are keen to avoid 'thinking under the null', for reasons I understand but do not completely share. If we view models as competing descriptions of the underlying population, then viewing the world through spectacles which in other language would be expressed as a null hypothesis is not necessarily a completely unnatural concept.

The fact that I would like to transfer the authors' ideas to other arenas serves simply to underline the fact that I found the paper very stimulating and I would like to repeat my thanks to the authors for presenting it.

Jim Ridgway (*Durham University*)

Statistical literacy should be about heuristics, not cook books

The paper paves the way not just for a radical rethink of statistics curricula, but also for approaches to statistical literacy. A key idea in the paper is that statistical novices can use heuristics which lead to judgements that a professional statistician would agree with, without necessarily understanding technical aspects of inference. Here, I argue that this approach should underpin efforts to increase public statistical literacy, and I offer an example.

In the UK, great emphasis is placed on evidence-informed policy. Examination results are used to create school league tables; schools are required to present student performance data in particular ways and are strongly encouraged to use these data to reflect on issues of teaching quality. The National Strategies Web site (`http://nationalstrategies.standards.dcsf.gov.uk/node/188855`) offers

guidance and instruction on how this can be done, illustrated by a school case history. Here is an extract, based on reading score tests from 10-year-old children:

'*Analysis showed that those making the best progress were:*

- *Girls...*
- *Middle to high attainers*'.

This is then qualified by 'because of pupil mobility, data available [*sic*] only for 15 out of the 24 who are currently in the year group'.

Statistical literacy is knowing that conclusions from such data are silly. An important mission for the Royal Statistical Society's 'getstats' campaign should be to ensure that government Web sites designed for key user groups should offer robust heuristics for interpreting data—such as the milestones in this paper. Milestones customized to particular sample sizes would be easy to generate on line. (Heuristics at a coarser grain size, e.g. *high dropout makes interpretation difficult*, could also be useful.) The current 'library' on the National Strategies Web site that provides clear explanations of important statistical ideas (regression, confidence intervals etc.) might better serve the intended audience if it focused on building statistical literacy via engagement with interactive displays.

Primary schools typically have year groups of around 30 pupils. Education effects are rather weak (Hattie, 2009). Wild and his colleagues show very clearly that conclusions about weak effects based on small samples are unwarranted. Primary school teachers need to be statistically literate. It is wrong to engage them in activities with high stake implications for their schools and professional activities that are conceptually flawed. This paper offers ideas and tools that are directly relevant to the statistical education of teachers.

John Pullinger (*Royal Statistical Society 'getstats' Campaign Board, London*)

Let me first congratulate Chris Wild and his fellow authors for the energy and inspiration in this paper. It is a rallying cry for our profession. We must find ways to make the ideas that underpin statistical thinking accessible to wider audiences if the discipline of statistics is to thrive and be appreciated for what it is: an essential life skill in today's world.

All too often we put people off. We seem dry and dusty. We make statistics look difficult. In schools we allow ourselves to be bit part players in mathematics lessons. Or, often even worse, we allow statistics to be mainstreamed in substantive subjects where students are encouraged to press a few buttons on a package and magic numbers pop out. We should never accept a null hypothesis. This paper argues with vigour that we need least of all to accept this nullest one.

The alternative is to make statistics interesting. The wholly visual approach that is advocated in the paper is appealing. The links that are provided in the paper take us to a series of mesmerizing pulsating boxplots. The approach makes the necessary intuitive connection to the fundamental ideas of statistical inference. I buy the analysis but I am compelled to ask for more.

Success needs to be judged, as in all fields of enlightenment, by the ability of the method to equip the seeker with a desire to apply their new-found skill in their own lives. They must have enough thirst to learn more on their own. To do this, the paper's rallying cry must call the profession to show just how pervasive statistical thinking is to everyday life and be its advocates. Each of us can show how many of the judgements that are made by us and those around us are based on a sample drawn from the fullness of reality. Each of us can show those we meet how to be enquiring about how those samples are drawn in order to make sound inferences for their choices. Each of us can help others, in schools, in work, in government, in the media and in everyday life to 'getstats'.

Thank you for a paper that calls us to action.

James Nicholson (*Durham University*)

The authors acknowledge that the detail of the paper is looking at a narrow area of the curriculum and a small part of the statistical process. I agree that there are very interesting uses of technology, especially in visualization, which may help to build much better understanding of significance where random sampling is involved. The paper also talks about wider issues in statistical education and it is to those that I wish to address the bulk of my comments.

One striking feature of the New Zealand curriculum is the early introduction of multivariate data, and the repeated revisiting of reasoning from evidence in gradually more sophisticated contexts. Teaching mathematics in the UK I have seen students who are mathematically talented, or at least very competent, feel that they were not needed by the social sciences in that they never encountered real social problems

as substantial contexts, nor worked with data where mathematics offered real insights into social issues. Those students with a leaning towards the social sciences, or wanting to address social issues, find the academic work mostly based on words with some headline statistics, usually of aggregated data, but little to suggest that there is a need for the students to be able to interpret the data for themselves, or that mathematics and statistics are potentially valuable assets.

As a consequence, in the UK at least, the current crop of social science students has a level of mathematical competency and statistical literacy which causes concern. The issues of dealing with multivariate data on the scale that is encountered in social sciences does not require statistical inference in the sense of making inferences about populations from random samples of quite small size. However, there is the task of reasoning from the evidence of large-scale multivariate data where headline statistics can be disaggregated by one or more factors to explore the nature of differences. For example, if on average women live longer than men, is this consistent across all ethnic groups?; across social class?; across all regions of the UK?; and is the situation changing over time? This is quite a different sort of statistical inference, but one which might have a broader usefulness than that addressed in detail in the paper.

I wonder whether the authors feel that the early engagement with multivariate data will do anything to reduce the perception which students seem to acquire that social sciences and mathematics belong in non-overlapping universes.

Thomas King (*University of Southampton*) **and Clare M. Woodford** (*Queen Mary University, London*)

There is a need for students to be taught more of the problem solving process and creativity of statistics, and also for these ideas to be accessible to all students. However, this leaves the questions of what is actually learned; what are the outcomes for different students; how can they apply this in their future life; and, how does this foster statistical literacy? If we define statistical literacy as being 'familiarity with reading and writing data stories' (King, 2011), then we need to think of how this may be developed as concurrent work shows this is not an immediate outcome (Pfannkuch *et al*., 2010).

The focus on visual thinking initially caused some concern as facility with different modes of thought varies substantially between students. However, the idea in the paper is to keep eyes on the graph whereas visual thinking typically means more abstract visualization than concrete comparisons (Grandin, 2006). To tell data stories visually, and also to assess students' competence, they can certainly present their own arguments, giving a commentary supported by their own graphs. Indeed such a presentation would engage other students in interpreting the data stories of their peers and reading their graphs. This would also bring a greater understanding of the difficulty of engaging with unfamiliar graphs which have typically not been understood by scientists (Roth, 2004).

In England particularly, students follow different paths at age 16 years. The Advisory Council on Mathematics Education (2010) proposes that from 2016 all students should follow some mathematics curriculum until age 18 years. This

> 'would help students develop mathematically as citizens [through] understanding and using statistics (including a critical appreciation of how they are presented and interpreted in the media)'

(Advisory Council on Mathematics Education (2010), page 5) even on the least technical level 3 pathway. Smith (2004) recommends that statistics should be taught outside mathematics, although this was not recommended by the government, perhaps partly because of teaching capability. Thus this paper would present the technical part of a wider implementation of statistics in the school curriculum but does not address the range of mathematical outcomes that students within their programme may have. Furthermore, if this will be such a new development in schools, how will the learning be supported by parents and teachers in other subjects, and does this strategy for developing inference enable conceptions of statistics as a versatile method for empirical discovery rather than a recipe for divining significance (see Petocz and Reid (2010))?

Julian Stander and Rana Moyeed (*University of Plymouth*)

We congratulate the authors on their educationally important and thought-provoking paper. We very much like the idea of producing a mental picture of superimposed or vibrating boxplots whenever we deal with data from a continuous random variable such as those shown in Fig. 5 and Fig. 6. This idea has already been included in an introductory statistics module delivered to a large group of students at the University of Plymouth as we believe that it will provide a valuable aid to in-class discussion of repeated sampling. Nevertheless, we contend that notions like repeated sampling from a parent population, sampling variation, confidence intervals, statistical significance and p-values will continue to present many pedagogical difficulties. Although we understand the need for students to have a good knowledge

of frequentist statistics, we would like to see the statistics education community give more emphasis to the development of tools for the teaching of Bayesian techniques, an excellent example of which is Tony O'Hagan's First Bayes software that is available from `http://www.firstbayes.co.uk`. Such tools can also enhance and enrich the teaching of probability ideas, a firm grasp of which is fundamental for the understanding of statistical inference. We shall shortly start to investigate with the Royal Statistical Society Centre for Statistical Education how visualization techniques motivated by those in this paper can be applied in the Bayesian framework. We believe that adopting the Bayesian approach to learning about underlying processes would help educators to present inferential ideas in a much more convincing and scientifically natural way, avoiding the issues and topics related to repeated sampling that are often so difficult to understand and explain. In any case, the authors' use in Section 3.3 of the phrase "'I'm thinking that the true median is likely to be in here somewhere'" suggests that people instinctively adopt a Bayesian approach for statistical inference.

Ramesh Kapadia (*Institute of Education, London*)
It is fascinating to hear a paper about teaching conceptions of statistical inference read in London three decades after the first major international project on teaching statistics completed its main work in schools in England (Schools Council Project on Statistical Education, 1980). As the project officer under Peter Holmes I fondly remember the innovative ideas expounded, many of which are now part of the English national curriculum.

The units of work that were produced for schools in the late 1970s did include some inferential work, such as on drawing implications from samples to populations. Yet, since the work was aimed at the 'bulk of students and not just an intellectual *élite*' (page 5), we avoided the definition given here of statistical inference as 'confidence intervals ... and ... distributions'. We encouraged pupils to look at data in context, without applying formal tests. We also aimed to underpin the work with the sound foundations that are afforded by probability. This aspect seems to have been underplayed in the current paper, but it is explored in Kapadia and Borovcnik (1991) and Borovcnik and Kapadia (2009). We would be interested to hear of how even the majority of students can really cope with ideas of formal hypothesis testing or milestone 3, since many pupils find probabilistic ideas difficult (Kahneman *et al.*, l982).

This paper uses visual possibilities of technology to widen the possible approaches: evidence on how well this works is now needed, notwithstanding the pioneering work of Hans Rosling. We would like to see detailed evaluative research testing the efficacy of such an approach. This needs to confront the relative lack of impact of statistical software and computers in mathematics education as a whole, in comparison with the sometimes quite extravagant claims that are made for geometrical and statistical packages.

The paper advocates that the required inferential steps should not be demanding and supports a rule-based approach. This ignores the fact that the process of reaching rules is pedagogically important and too camouflaged here.

We would also be interested in reading about national assessment from New Zealand where it is '*impossible*' to obtain a good mark by blindly applying tests and high marks will require 'understanding and insight'. Even a few examples of such assessment would mean that New Zealand led the world, for such questions are all too rare in examinations in England.

The following contributions were received in writing after the meeting.

Alan Agresti (*University of Florida, Gainesville*)
I commend the authors for making sensible proposals for improving the teaching of statistical inference at various stages of the education process. I like their emphasis on a gradual building of complexity using visual displays of sampling variation.

Like the authors, I believe that the inference part of most introductory statistics courses overemphasizes significance testing. I agree that it is more informative to focus on answering 'Which is bigger, and how much bigger?' than 'Are they different?'. We all recognize how much difficulty students have in 'thinking under the null' and properly interpreting p-values. Given its difficulty and how few of the students at this level will ultimately be conducting research or reading research journal articles, I do not think it warrants so much attention. In addition, for a staged development path in schools such as the authors recommend, many (probably most) of the teachers themselves are unlikely to understand well the subtleties of significance testing. By contrast, I think that students find it easier to understand a confidence interval than the many steps of a significance test. In the authors' milestone 4 introducing formal inference, I would also suggest explanation of confidence intervals and the distinction between practical and statistical significance.

The authors' milestones emphasize the analysis of quantitative variables, treating categorical variables in an extension. I believe that the authors' complexity reduction strategies in explaining sampling variation would be enhanced by also discussing categorical data at early stages, focusing mainly on the binary case. The students will already have heard about polls and other surveys, so it is easier for them to comprehend variability from sample to sample in a percentage than variability in a boxplot. Simulation to illustrate sampling variation is also simpler to explain with binary data. At an early stage, the instructor could connect a binary version of the Fig. 11 footprint with the sample size by noting that the simple error bound of $100/\sqrt{n}$ (on the percentage scale) describes variability of the footprint.

In summary, our students' understanding and appreciation of the importance of statistics would improve, and our profession would ultimately be better off, if more of us followed the authors' lead in thinking seriously about these issues and then making concrete proposals for change.

Janet Ainley (*University of Leicester*) **and Dave Pratt** (*Institute of Education, London*)

The paper presents an innovative pedagogic sequence for leading students towards an understanding of sampling variation. It is argued that current simulation methods involving resampling confuse students. Naive learners' grasp of the key concept is tenuous and so it is reasonable to conjecture that approaches which follow the heuristic 'keep the eyes on the graph' might focus attention. We look forward to this idea being tested.

The approach proposed aims to support a relational understanding of sampling variation by employing an analogy in which the actual patterns in the population are distorted through sampling variation. Current pedagogic methods continue to emphasize an instrumental approach in which procedures are temporarily learned and soon forgotten because of their meaninglessness.

A continued tension, however, is that, although the pedagogic focus is on gaining a relational understanding of sampling variation, students are likely to lose any sense of how this notion could be significant for them. It is one thing to understand the difference between the description game and the inference game, but that matters little if they see no point in playing the game at all! We therefore advocate an early emphasis on investigations that throw up the need for sampling and issues about why sampling variation is useful. In other words, we want milestone 1(a) to be 'writ large' so that a utility-based understanding might be constructed and not forgotten in the imaginative learning trajectory that might afford a relational understanding.

Such tasks would need to focus on the question 'If you are trying to answer a question about the population, why look at a sample?'. In a real life situation, the answer to this question is that the population is not accessible, because it is very large (in the case of market research or opinion polls), because it is not finite (as might be the case in a scientific experiment) or it is impractical (as in a blood sample). An example in this paper is a sample taken from the CensusAtSchool data for New Zealand. Taking samples from this is a pedagogic device, since all the data in the population are available. Such pedagogic artifices are often deployed when relational understanding is sought but we need to remember that, in such situations, the lack of authenticity works against utility-based understanding.

Murray Aitkin (*University of Melbourne*)

As I have not taught statistics at school level my comments are relevant to the university introductory course for non-mathematicians.

I have had happy success with the approach of Hodges *et al.* (1975). The finite population approach in which students draw their own random samples by dice throwing seems to me the essential place to begin: it raises randomness concepts immediately, which can be followed up with the examples of voluntary sampling biases in Moore's several books, and clinical trial design. It also allows us to define probability, which is a term that is apparently missing from the early courses described by the authors, in terms of equally likely cases based on the die regularity.

I restricted the inferential content to population proportions. Once we have probability, it is straightforward to compute probabilities of events, and now Bayes theorem for simple two-alternative hypotheses becomes simple to use, also replacing Venn diagrams by two-way contingency tables.

To make inferential statements by plausible intervals we can expand the two-alternative case to the K-alternative case—the advantage again of finite populations is that the probabilities are all small but finite, and the usual credible interval calculation can be introduced as an approximation to the real finite sum. For continuous variables we dichotomized them at the median.

In the course I last gave at the University of Newcastle, I interspersed the discussion of probability and sampling with set piece lectures on specific real world studies which illustrated the importance of proper design—the analysis was not the issue.

I had the highest approval rating ever from students in this course, and I enjoyed it more than any other course I have taught. The weakest student said to me that she really enjoyed it and came away from it with the conviction that statistics was very important in all areas of society.

It was a revelation to her that her 95% interval did not include the true value, though those of all the other students did! She asked with great embarrassment 'What did I do wrong in my sampling?'. 'Nothing!' was my reassuring answer. 'You are discovering an inherent limitation of statistical inference—we can only make 100% confident statements which are trivial!'

A freely downloadable book for this course can be found at `http://www.ms.unimelb.edu.au/~maitkin/InfModWorld.pdf`.

Adrian Baddeley (*Commonwealth Scientific and Industrial Research Organisation, Floreat, and University of Western Australia, Perth*)

First I wish to emphasize the importance of this topic in the widest sense. It is already clear that some of the major scientific errors and technological failures of the 21st century will be caused by erroneous inferences from data. An increasing proportion of research effort in many disciplines is devoted to data analysis rather than laboratory or fieldwork. New research findings rest increasingly on inferences from data, so it is vital that the key ideas of statistical inference be absorbed into general knowledge, It is now crucially important for our discipline to take seriously the task of communicating statistical ideas, particularly statistical inference, to a wide audience.

Arguably the authors' project is not just about finding new and better ways to communicate key statistical concepts. Perhaps on the principle that we do not really understand something until we can explain it to a child, the project is also a fundamental investigation of the logical structure of statistical inference. The challenge of presenting the key ideas in graphical form is also an investigation of connections between statistics and geometry.

The authors demonstrate some of the wonderful opportunities that are presented by new information technology for communicating statistical concepts. This technology also presents our greatest threat, namely the 'drag-and-drop' model of data analysis, the growing expectation that statistical analysis can be reduced to a software process. To respond, we need to identify and emphasize those parts of statistical methodology which truly cannot be automated, and this again is related to the authors' project.

First-year university students in microbiology are shown how to extract DNA from a tissue sample in their first laboratory class. They follow a recipe and are completely unaware of the complex biochemical and physical processes that are involved; yet the excitement in the room is palpable. Our challenge is to match that level of excitement when we teach statistics.

Manfred Borovcnik (*University of Klagenfurt*)

The authors' approach to statistical inference is visually oriented and rule based, but not closely linked to context (which is not discussed here), decision logic and probability. The comparison of two distributions is related to the key question of 'When can I make the call that B tends to give larger values than A?'(pages 259 and 270), which lacks an easy answer and might lead one astray. The probability of a paired comparison being incorrect can be high for various distributions.

The visualization of sampling fluctuation is compiled by the superposition of repeated boxplots. However, the blurred graph may emerge from boxes that are centred at the same point widening or narrowing, or, with the same width shifting left and right. More importantly the *size* of fluctuation is caught only *qualitatively*. Mathematically the minimum and maximum (marked by whiskers) of larger samples are spread more to the extremes; yet, paradoxically, the diagrams (page 258) show the opposite.

The three suggested milestones vary in the inherent α-levels of the rules: it varies in milestone 1 from 15% to 0.4% (with n) and is about 8% in milestone 2, and 2% in milestone 3. So more sophisticated milestones, perhaps perversely, do not necessarily lead to better outcomes, depending on the sample size: moreover, the power to detect a difference decreases by later milestones, which is counterintuitive statistically and perhaps misleading.

There are situations where a difference is correctly detected by the crude rule of milestone 1 but not with the rule of milestones 2 or 3. For a difference in the means of two distributions of the size of 0.6745σ (with common standard deviation σ), rule 1 (of milestone 1) has a power of 0.66, and rule 2 only of 0.52, whereas rule 3 goes down to even 0.44 (as estimated from a simulation study from normal distributions with 30 data).

The poorer behaviour of the milestone 3 rule is partly caused by the 'intuitive intervals', which treat extreme and usual cases as equally likely. This procedure is defended by the authors but makes the 'confi-

dence' interval broader by a factor of roughly $\frac{1}{3}$ and precludes a probabilistic argument. Thus this is not a good progression to the more conventional milestone 4 of formal statistical inference.

Besides the inconsistencies noted, the visual representation helps only to *memorize* the rules, but not to understand them. Where is the cognitive development as rules are not motivated?

Mike Camden (*Statistics New Zealand, Wellington*)

This contribution aims to raise two issues. The first is that a widespread understanding of inference is vitally important for the quality of official statistics. The second is that an understanding of sampling uncertainty for estimates is valuable in itself. The methods in this paper are used mainly for comparisons, but they could be adapted for estimates like medians and proportions.

The authors ask 'why should statisticians care ...' about school statistics? Official statistics agencies have good reasons to care. Much of the agencies' efforts go into sample surveys. The quality of the results from these surveys depends on response rates. The approaches in this paper will enable current students and future citizens to understand the aims of statistical inference, and so to value their contributions to social and business sample surveys. A statistically informed society and high response rates are vital for minimizing sampling error as well as bias.

The authors make the results of sampling visible for estimates of centres (Figs 5 and 6) and proportions (Figs 11 and 12). However, much of their thinking centres on comparisons, and 'making the call' about differences. A task for the future may be to apply their thinking to the estimates, and indeed to estimates of differences. Questions about estimates may be less motivating than questions about comparisons, but they are common in official statistics, and in the rest of the work of 'data detectives'. As the authors note, the vibrating boxplots of sample medians lead to intervals of uncertainty. Perhaps these can be used by students to make informal inference statements about population medians and similar measures.

The authors note that 'computer technology has changed the world entirely'. Methods like those in this paper, soundly designed in terms of both statistics and pedagogy, will enable today's students and tomorrow's citizens to contribute to and to use data wisely. Perhaps they will say 'I helped to reduce the uncertainty in that estimate', and be proud of it!

Len Cook (*Victoria University of Wellington*)

The on-going revolution in information and communications technologies brings new challenges to the numeracy of populations in all countries. How to make this a compelling challenge to people at all stages of learning is of serious concern to which statisticians have generally not yet responded successfully.

The ideas and methods in this paper are a significant addition to the statistical toolbox that we need for our role as educators in statistical reasoning. They extend the capacity to lift numeracy, not only of those involved in teaching, but also of those who develop and present evidence.

The societies that we live in are in the midst of responding to major transitions, and common to almost all are the effects of social and demographic change, globalization and climate change and information and communications technology. We can effectively plan the future form of but few activities and systems without a rich understanding of the context in which they will operate, and the understanding that we need will come from understanding and evaluating evidence from both traditional and innovative sources of information. Most decisions will be taken by those without a trained scientific understanding of the validity of the evidence gained from traditional approaches, whatever their field of endeavour.

Developing simple means of bringing concepts of variation and statistical inference to a wide range of decision makers and analysts, whatever their stage of learning, would be an extraordinary achievement, in terms of its effect on the quality of decision making in such uncertain times. This paper is also a reminder that those involved in the development of evidence, and its evaluation, have obligations to ensure its accessibility not only to those involved in policy development and advice, but also for informing the wider community about whatever new knowledge we have. We are reminded that this includes ways of assessing evidence, as well as knowledge of the evidence itself. To this extent, the authors should have as their target those who engage in scientific activity in all fields, simply because public understanding in an informed democracy of the evidence that is produced about communities, societies and the planet itself is vital for the political choices that we all now face.

Neville Davies (*Royal Statistical Society Centre for Statistical Education, Plymouth*)

Scanning the Society's journal over the last 173 years it is clear that there are few papers that address the teaching of statistics, let alone help beginning learners of it. Therefore I particularly welcome the fact that this paper is designed to help teaching *and* learning in an area that some regard as 'obvious' or 'simple':

inference is often difficult for newcomers—especially non-specialists. It is all too easy not to realize the difficulty that students have in getting to grips with the concepts of inference.

The authors brilliantly marry easy-to-use technology with exploitation of visual senses to produce better understanding of the relationship between samples, their sizes, variation, sampling and how these relate to what are 'back in the populations' (adopting the authors' catch-phrase). These result in simple but giant steps forward in helping to teach *and* learn these aspects of statistics. However, I believe that the authors should stress *more* that the approach is useful for beginning learners of inference at *any* age.

The authors make extensive use of real survey-type data from *CensusAtSchool* New Zealand, one of several similar projects run in six other countries, including the UK. The data produced have been useful to help to develop statistical thinking at school and other levels (Connor *et al.,* 2006). The databases from each country's implementation are stored at the Royal Statistical Society Centre for Statistical Education (RSSCSE) and contain well over 1.4 million responses that can be sampled over the Internet. The RSSCSE intends to build from the ideas in the paper, exploit the family of databases, augment them with others that reflect data production methods from designed experiments and use technology and visualization tools to create an electronic Wiki-type dynamic collaborative teaching, learning, assessment and continuing professional development environment. It will be free to use and will support the RSS 10-year statistical literacy campaign (`www.getstats.org.uk`). There will be six dynamically connected sections:

(a) data visualization, interrogation and presentation using the authors' diagrams with memory and geographical information systems;
(b) investigation and problem solving, including opportunities for discussion and review;
(c) teaching resources with a shareable teacher's corner enabling exchange of dynamic documents;
(d) learning resources, including a learner's corner with self-reflective journals;
(e) embedded assessment, including formative, summative and a problem solving approach to teaching and learning (Marriott *et al.*, 2009);
(f) continuing professional development and access to on-line resources leading to accreditation by the RSS.

The environment will be available on the RSSCSE Web site in due course.

N. I. Fisher (*University of Sydney and ValueMetrics Australia, Sydney*)

I congratulate the speakers on their efforts to interest the wider statistical community in what—and how—school students learn about statistics. And New Zealand is to be congratulated on its leadership in establishing a national curriculum labelled *mathematics and statistics*.

Their paper identifies difficult challenges in introducing a school statistics curriculum that will, on the one hand, equip all students with some life skills which they will need in dealing with the vagaries of life and, on the other, attract some of the abler students to pursue further statistical studies.

Currently, in Australia, statistics is trapped in a game of double jeopardy, incorporated invisibly in a curriculum area called mathematics (seemingly its most obvious home); and, as a consequence, mathematicians believing that ownership of statistics is rightfully theirs. Since 2002, the leadership of the Australian Bureau of Statistics and of the Statistical Society of Australia have been working on strategies to revolutionize statistics education from kindergarten to post-doctoral research. It has been an arduous journey, with occasional moments of great promise followed by lengthy periods of desolation (Fisher, 2010). We are now close to having a statistics stream in a new national K-12 Mathematics (*sic*) Curriculum, but we are still battling to have it read like a statistics curriculum developed by statisticians and not by mathematicians. (However, there are grounds for hope: after one lengthy discussion, a mathematics educator finally remarked to me, 'Ah, I get it. Statistics starts with a question'.) Many mathematics teachers do not appreciate that

> '... Statistics is a science ... and it is no more a branch of mathematics than are physics, chemistry and economics; for if its methods fail the test of experience—not the test of logic—they are discarded'

(Tukey (1953), quoted by Brillinger (2002)). Much of statistics can be taught without formulae, a partial corollary to Efron and Tibshirani's (1993) observation that

> 'The traditional road to statistical knowledge is blocked, for most, by a formidable wall of mathematics'.

There have been many interesting lessons to learn from our experience.

Perhaps statistics education and the related area of professional accreditation should be viewed as part of public awareness, and so benefit from a broad, professionally developed strategy based on synthesizing the numerous, very worthy, yet disparate activities (booklets, e.g. *Statistics: a Job for Professionals*, available

from `http://www.statsoc.org.au/statistics.htm`, Web sites such as `http://understandinguncertainty.org/`, journals like *Significance*, ...) that are currently in place.

Joan Garfield and Andrew Zieffler (*University of Minnesota, Minneapolis*)

Information and formal statistical inference?: new questions raised

Wild and colleagues have done a laudable job of describing problems that students have in learning to make statistical inferences and in also creating an innovative approach to prepare school children to reason informally about statistical inference. The ideas espoused in the paper, with the pedagogical focus at earlier grades on the big concepts underlying inference (e.g. sampling variation), rather than obfuscation via procedures (e.g. *t*-tests) and theory (e.g. the central limit theorem), are at the heart of informal inferential reasoning.

Informal inferential reasoning is defined by Zieffler *et al.* (2008) as

> 'the way in which students use their informal statistical knowledge to make arguments for connections between observed samples and unknown populations',

including

(a) reasoning about possible characteristics of a population (e.g. shape and centre) based on a sample of data,
(b) reasoning about possible differences between two populations based on observed differences between two samples of data (i.e. are differences due to an effect as opposed to just due to chance?) and
(c) reasoning about whether or not a particular sample of data (and summary statistic) is likely (or surprising) given a particular expectation or claim.

Informal inferential reasoning has been the focus of recent research and discussion, as it appears to offer a promising way to help students to build important concepts of statistical inference. Furthermore, it allows the study of how students intuitively reason about sophisticated ideas that are usually difficult to formalize and use. In Zieffler *et al.* (2008) we suggested several questions that emerge from the study of informal and formal statistical inference that can guide research. The approach presented by Wild and his colleagues suggests many other questions, such as the following.

(a) What are effective ways to develop students' informal inferential reasoning?
(b) How does good informal inferential reasoning foster students' ability to use and understand formal methods of statistics inference?
(c) What aspects of formal inference are needed given current tools and approaches (e.g. randomization methods)?

Wild and his colleagues have created an approach worth studying that can help to address the first two questions. However, a key question is whether good informal inferential reasoning can stand alone, or whether it must be a stepping-stone to formal methods of inference. If students develop good ways of reasoning about the uncertainty of conclusions drawn from sample data in light of variability, would that not be a worthy end goal of statistical instruction?

Andrew Gelman (*Columbia University, New York*)

I agree that, wonderful as informal plots and data summaries are, we also should be teaching students formal statistical inference, which is a big part of what separates statistical thinking from mere intelligent commentary. I like the authors' formulation that statistical inference

> 'addresses a particular type of uncertainty, namely that caused by having data from random samples rather than having complete knowledge of entire populations, processes or distributions'.

The authors write

> 'we also need to start from a clear distinction between when we are playing the description game and when we are playing the inference game'.

I would go one step further and get rid of the concept of 'description' entirely, for two reasons.

(a) Ultimately, we are almost always interested in inference. Catch-phrases such as 'let the data speak' and rhetoric about avoiding assumptions (not in the paper under discussion, but elsewhere in the statistics literature) can obscure the truth that we care about what is happening in the population; the sample we happen to select is just a means to this end.

(b) Description can often—always?—be reframed as inference. For example, we talk about the mean and standard deviation of the data, or the mean and standard deviation of the population. But the former can be presented simply as an estimate of the latter. I prefer to start with the idea of the population mean and standard deviation, then introduce the sample quantities as estimates.

Similarly, some textbooks first introduce linear regression as a data summary and then return to it later in the context of statistical inference. I prefer to start with the linear model $y = a + bx$, with no 'ε'—I believe it is a mistake to focus on the error term before students have become comfortable with the deterministic model—and then introduce the least squares estimate, not as a data summary, but as an estimate of an underlying pattern of interest.

In giving these comments, I am not trying to imply that my approach is the best way or even a good way to teach statistics. I have no evidence to back up my hunches on how to teach. But I would like to suggest the possibilities, because I think that statisticians are so stuck in a 'description *versus* inference' view of the world which can lead to difficulty in teaching and learning.

Harvey Goldstein (*University of Bristol*)

I very much enjoyed reading this paper, and especially its emphasis on the progressive refinement of statistical understandings. One of the issues that is not discussed by the authors, apart from a brief mention in connection with Fig. 12(a), is that of dependences between data values. In all their other examples is embedded the unarticulated assumption that data values have been independently sampled and there is also, throughout most of the examples, the assumption that the sampling procedure is 'unbiased'.

It could, of course, be argued, as the authors do in connection with Fig. 12(a), that these are complexities that can be introduced at later stages. This assumes, however, that in an important sense the notions of unbiased and independent sampling are more 'basic' or more 'natural'. Although they may be considered formally more basic they may not be so pedagogically. Thus, in practice, samples are often biased—e.g. through non-response—and also may contain (subtle) dependences, such as occur in multistage sampling or where time series are involved such as in successive daily temperature measurements. If such dependences or biases are important then the early milestone inferences may not be even approximately correct. The provenance of the data is relevant and important. Thus, the example of pupil heights that was discussed by the authors implicitly assumes both independence and unbiasedness, since the authors are concerned with making inferences to a real population—although what the population consists of is not actually described. When the authors discuss the computer animation of sampling variation they are (I assume) generating simple random samples from a 'population'. But these do not necessarily have the same statistical properties as real life samples—so how are students to be made aware of this artificiality?

My suggestion is to introduce ideas of bias and dependence early on, say in milestone 1. I take the point that one does not wish to introduce too many concepts at once, but both of these ideas are so fundamental to all inference that to ignore them runs the danger of divorcing the pedagogy from the reality. This issue is generally not a problem in traditional 'formal' statistics courses since the simplifying assumptions of unbiased and independent sampling are explicit, later to be modified. The challenge in the curriculum that is proposed by the authors is how to deal with them at the outset in ways that are intuitively accessible to students. I suspect that to do this satisfactorily would involve the extensive use of real data before, or alongside, artificially generated data and I would very much welcome the authors' comments.

Robert Gould (*University of California, Los Angeles*)

The authors are to be congratulated on developing an innovative and exciting new curriculum for teaching statistical thinking to young students. Most exciting to me is the authors' conception of students as lifelong learners who will 'live statistically' and who therefore need to internalize statistical thinking. In the past, students first encountered data in their introductory statistics course, where they were taught formal data collection methods and warned that inferential techniques applied to only very special circumstances. Course content focused on producing critical consumers of statistics. However, modern students encounter data routinely, in both formal and informal environments, and are far more likely than prior generations to need to process and analyse data. The need to produce citizen statisticians has never been greater.

This demand to produce statistically literate citizens has perhaps caused many instructors to overload the curriculum. Wild and his colleagues take a different tack and strip away all but the essentials. This decision to focus on inference in a limited context at such an early age will lay a strong foundation for learning some of the more complex statistical concepts at a later stage. Even those of us who must teach 'all of statistics' in one (too short) course could benefit from the example of choosing a small number of topics and teaching them very well, rather than lightly touching on many topics and teaching very little.

One lesson learned from this paper is that the proper use of technology can strip much of the tediousness from the curriculum and allow students and instructors to focus on important conceptual features.

As Hotelling (1948) explained in his historic summary of the then-current state of statistics education, *who* teaches statistics is as important as what is taught. The curriculum proposed can be taught only by teachers with a firm conceptual understanding of inference and data analysis. This is not meant as a criticism of the authors' work, since even the best curriculum can be ruined by a poor teacher, but instead as a reminder that, in most classrooms, statistics education is left in the hands of people who have not themselves learned statistics. Still, to prevent the lessons from being anything more than a transmission of a set of rules (an approach which has dominated statistics education), improved professional development will be required to ensure that teachers of the curriculum proposed will have a sufficiently deep understanding of statistics.

Sander Greenland (*University of California, Los Angeles*)

I am disturbed by the authors' exclusive focus on random variability. In evidence synthesis and observational research, random variation can be smaller than uncertainty about bias (Greenland, 2005; Lash *et al.*, 2009). Even in randomized trials, extensive drop-out and non-adherence can render naive any claim to resembling the experimental designs on which conventional statistics are based. Sample surveys can suffer similarly through non-response and erroneous response.

These issues are central (not tangential) to sound understanding and application of statistics. Formal statistical inference extends to non-random sources of variation and uncertainty (e.g. Gustafson *et al.* (2001), Vansteelandt *et al.* (2006), Turner *et al.* (2009), Molitor *et al.* (2009), Greenland (2009) and Geneletti (2009)). Yet Section 2.1 of the current paper says that statistical inference

> 'addresses a particular type of uncertainty, namely that caused by having data from random samples rather than having complete knowledge of entire populations.... It does not address study design and execution issues, quality of data, relevance of data, practical importance and so on.'

I think this conceptualization is harmful, making statistics a bag of blind algorithms rather than an integral part of *scientific* inference. Worse, it seems to license the usual application to highly non-random data of algorithms that assume purely random variation.

Simply put,

> 'it is misleading to report confidence intervals which reflect only the extent of possible random error when systematic biases are suspected'

(Turner *et al.*, 2009). Despite such *caveats*, basic statistics courses often omit serious discussion about the non-random problems of real studies. Consequently, researchers rely on quantifying only random variation, and tend to underestimate seriously the uncertainty that they should associate with a given result. They may see a confidence interval which summarizes only random variability, and then *behave* as if that captures most uncertainty (even when claiming to know better). I fear that the authors' approach will only further propagate and entrench such overconfident behaviour, making it even more difficult to unlearn than it is now.

Fortunately, one can introduce bias concepts in tandem with random variation without leaving the authors' framework or introducing untoward complications. For example, one could illustrate what might happen in a survey in which heights are asked of the children rather than measured directly, when there is a tendency towards under-reporting by girls and over-reporting by boys. Showing how such problems cast doubt on conventional rules (like those given by the authors) would be a worthwhile educational achievement; it should not be left to a separate module.

Paul Hewson (*University of Plymouth*)

I find the paper very interesting and the underlying body of work very encouraging. I thank the authors for that. I have two questions: one specific; one general. The specific question concerns the authors' statement in Section 2.5 that the 'experimental *versus* observational' issue is a tangential concern. I am not really sure that I entirely follow this point. I wonder whether this dichotomy is so fundamental that the authors have really presented a method for engaging with observationally based inference. One could develop a parallel, experimentally based pathway, which would have many stages in common but be fundamentally different precisely because the dichotomy is so fundamentally important.

I was encouraged to read that randomization-based inference will be introduced at a school level as this seems a more direct way of encouraging people to think under the null. But my more general point is that

I did note the authors' repeated comments that 'thinking under the null' should be delayed as much as possible. I think that this is a great idea, but I wonder why not delay it indefinitely? The Bayesian paradigm is widely accepted in applied statistics and practice nowadays. Additionally, noting that (to give a Waikato-based example) Bolstad (2007) has produced a highly acclaimed textbook which teaches non-statistical specialists how to work in a Bayesian framework, we have support for the view that we do not actually need any thinking under the null. I think that the general point also relates to the specific—thinking under the null seems simpler to introduce in the context of experimentally based inference than observationally based inference, and I therefore think that the proposals in this paper really lean towards Bayesian inference.

Kuldeep Kumar (*Bond University, Gold Coast*)
This is a very interesting paper dealing with teaching statistical inference to high school students and also as an explication for a first course at university level. Statistical inference is of course one of the difficult subjects to teach and I must congratulate the authors for providing insight into teaching this subject. In most of the business schools statistical inference is taught only as a part of the statistics for management course and I think that it is much more challenging to teach the concept of statistical inference to management students. One of the reasons is the low numeracy skills of the students and sometimes, secondly, difficulties arising in teaching mature age students who have left mathematics long ago. The Business School cannot afford to have a full course on statistical inference. Accordingly, the question is how can we teach statistical inference to these students? Especially in the context of a business school there should be more emphasis on teaching the application of statistical inference rather than the theory. There has been much interest in the teaching of statistics in business schools for a very long time; for example see Kumar (2010) and references included in the paper. We usually use problem-based learning where all the topics are introduced with a problem in real life. However, maybe in a business context, the statistical tools should be taught to show how the problem can be solved, and thus exploring the help of these tools and techniques by using real data.

Teaching basic statistics courses by using Microsoft EXCEL has become more popular in business schools. We also quite often use EXCEL to demonstrate the central limit theorem and other concepts dealing with statistical inference. One of the advantages of using EXCEL is that it is very easy to use and it comes as part of the Microsoft package which is available on most personal computers. EXCEL is a very good tool for graphical representation and producing tables. The 'Analysis toolpak' can do many statistical analyses, e.g. calculation of descriptive statistics, correlation, multiple regression, analysis of variance, the t-test, Z-test, confidence intervals, moving averages and the random-number generator. The basic statistical functions can be used to calculate normal probabilities, binomial probabilities, χ^2-tests etc.

Usually I have observed that school children are good at using EXCEL and I wonder whether the authors can use EXCEL to demonstrate the concepts that are mentioned in the paper in a more accessible way.

D. V. Lindley (*Minehead*)
The authors of this paper have several good ideas on *how* to teach statistical inference in first courses; but I have reservations about their ideas on *what* to teach. They rightly say

> 'Many of the problems with students learning statistics stem from too many concepts having to be operationalized almost simultaneously'

(Section 4), so why not teach *one* concept which will embrace the *whole* of worthwhile statistics? That concept is probability with its three rules of convexity, addition and multiplication. Probability can be taught by the physical experience of drawing balls from an urn, so that students can collect their own data.

Statistics is about uncertainty. The authors emphasize variation in repeated samples from a population, which is a source of uncertainty, but it is not the only one; what of the uncertainty in a court of law associated with a defendant's guilt? Uncertainty can only be satisfactorily expressed through probability, which is essentially an extension of standard deterministic logic to embrace uncertainty. The concept of a population is often artificial and rarely relevant to a practical problem. Repetition is not basic to statistics; coherence is and probability is coherent. The idea of a million apples falling from Newton's tree is not exciting but that the phenomenon coheres with the motion of the planets is. Boxplots are not basic but merely useful decorations, whereas probability is basic and not just pictorially effective. Even experienced statisticians can fail to understand the correct concept of a confidence interval: were the procedure to be carried out indefinitely in samples from the same population, then 95% of the resulting, intervals would include the true value. Often they think of the true value lying within the interval with 95% probability, which is typically false.

The authors' view of statistics is too narrow, restricting themselves to frequentist ideas, whereas operationally we need opinions that necessarily involve probability.

Thomas A. Louis (*Johns Hopkins Bloomberg School of Public Health, Baltimore*)
It is a pleasure to comment on this communication, especially on World Statistics Day. It will improve quantitative literacy, will add breadth and depth to statistical thinking and will enhance appreciation of statistics and statisticians. Students at all levels will benefit from the examples and viewpoints. I endorse the authors' call to action and most of their specifics, but I add my list of 'most important and fundamental concepts'.

(a) *The power of sampling*: the non-intuitive and impressive power of sampling to produce effective inferences from relatively small samples helps to make the case for statistics and statistical thinking.
(b) *The essential role of the sampling plan*: without information on the sampling plan it is usually impossible to relate observed data to the reference population. Simple random sampling is the place to start, but also introduce length-biased and size-biased sampling (the waiting time paradox is a great 'hook') and other sampling plans. There are a host of Bayes or frequentist issues related to this topic (see Mandel and Rinott (2009)).

A criticism: I am not in favour of using the term 'distortions' to characterize the effect of simple random sampling; it suggests chicanery. Why not use 'variation' in this context and save 'distortion' for other sampling forms?

(c) *The importance of strategic thinking*: for example, ask whether a conclusion should depend on whether distance is measured in miles or metres or a risk ratio is A/B or B/A. Use these to motivate use of (statistic − centre)/scale and logarithm-based procedures.
(d) *The fundamental role of the full probability distribution*: full distributional thinking promotes discussion of inferential goals. For example, it may be that the median or some other percentile is more relevant than the mean, or that the quantity of interest is the random variable induced by filtering a stochastic input (e.g. exposure) through an output function (e.g. an exposure–response relationship). In teaching the Gaussian distribution, show that the mean and standard deviation fully describe the distribution but are not necessarily the primary objects of inference. Use a similar approach for other parametric distributions. Full distribution thinking is essential for Bayesians (for whom it is innate) and frequentists.
(e) *The use of live data*: recruit students via interesting and timely data. For example, discuss issues in modelling of the spread of influenza, publication bias, stochastic properties of the yield from a screening programme, estimating the size of a crowd or civilian casualties, political polls, stock market gains (and losses!) ....

In summary, I recommend emphasizing the foregoing when educating, collaborating and communicating. I encourage all of us to build on the authors' work and thereby to promote statistics and statisticians.

Helen MacGillivray (*Queensland University of Technology, Brisbane*)
The deserved congratulations to the authors for the innovative, thoughtful and hard work of this paper should be matched by a determination of the statistical community to follow the examples set by them and others to become seriously involved in the many and diverse challenges in statistical education. There should be no illusions about the toughness of the challenges, the value of tackling them and the need for significant and scholarly work and analysis that bring together statistical and teaching expertise and experience.

Key points of this paper include the building of understanding gradually through revisiting and extending simple procedures, richer use of technology than merely reducing calculations, and that statistical soundness should underpin any simple procedures.

All statistical procedures need to be seen as part of what Chambers (1993) called 'greater' statistics. Emphasis on more holistic and practical approaches to statistics, reflecting what statisticians do, has played a crucial role in major developments in statistics education, including articulation and integration of the data-investigative process or cycle (Holmes, 1997; Wild and Pfannkuch, 1999).

Similarly, the skilful and integrated use of technology in statistics education should not prevent us from developing statistical thinking away from the computer. In discussing the teaching of sampling and sampling variation, the 'downward' view of sampling from a known 'population' or distribution can be matched by the 'upward' question of such importance in statistical inference, namely of what general

situation or group can we consider our data to be representative with respect to the question(s) of interest (Utts and Heckard (2007), page 72). Again, the data-investigative cycle and emphasis on real and everyday situations provide many opportunities for developing this awareness across all educational levels.

As in this paper, by choosing to illustrate simple procedures within the context of real many-variabled data sets, 'multivariatitis' can become the norm in teaching statistics. However, the tertiary level brings other questions, e.g. why not introduce statistical inference through categorical data? And should we not be using technology to investigate real many-variabled data sets sooner than in the traditional approach? Statisticians who help postgraduates across disciplines are aware of the consequences of overexposure to the traditional two-sample approach.

Perhaps we should develop 'worry questions' for teaching statistics, noting Bruner's (1996) point that most people have acquired a 'folk pedagogy' that reflects certain 'wired-in human tendencies and some deeply ingrained beliefs' (page 46). Asking questions, lateral thinking and analysing information are part of good statistical practice, including in teaching statistics.

Xiao-Li Meng (*Harvard University, Cambridge*)

'We don't care about school stuff.' Surely any professors who want good students (and who does not?) care, and even those who only care about research care about 'school stuff' in effect because they love to have well-taught students as research assistants! The issue, as the authors noted, is the concern that statistical education in high school can do more harm than good when there is a severe shortage of qualified statistical educators (Meng, 2009a, b, 2010a). The shortage is real, even at the university level (Meng, 2009c, 2010b). However, the authors' reasoning that we should start as early as possible is compelling, especially in this information age when everyone and everything are competing for the attention of future generations. The proposal to use modern technology is therefore appealing because technology can help, among many other things, to simplify and globalize pedagogical presentations, thereby reducing the dependence on local expertise for quality education, at least to some extent.

Regarding the authors' emphasis on concentrating on a single concept, arguments can go either way, as the authors noted. I cannot convince myself either way even after raising several glasses. Forcing beginners to dive in when they cannot swim is fatal. Allowing them to believe that they can jump into any wet environment because they have learned diving in a swimming pool is equally fatal. The right solution, I believe, is to start simply but also to demonstrate complications. It is safer to start with a swimming pool. This, however, should be supplemented by also showing beginners pictures or videos of brooks, creeks, streams, reservoirs, ponds, springs, estuaries, rivers, lakes, seas, oceans, or even waterfalls, wells and aquifers. They do not need to learn how to swim or dive in all these environments (nobody can!), but knowing about their existence helps to put what they have learned into perspective. After students have learned about sampling variation by using the authors' ingenious vibrating boxplots, why not 'vibrate' them more with a few relatable examples such as noise in digital transmissions, errors in forecasting the stock market and uncertainties in predicting their own examination scores? There is no need to spell out what causes these 'vibrations'. The mere recognition that they are real, need to be studied but do not fall into the category of 'sampling variation' should serve well the purpose of inspiring students to study statistics as a vast, vital and vibrant scientific discipline.

Deborah Nolan (*University of California, Berkeley*)

I congratulate Wild, Pfannkuch, Regan and Horton for a bold innovative approach to teaching statistical inference. They think critically about concepts in inference; they separate and scaffold the big ideas, and create a model for how technology can aid in understanding these concepts. Too often we allow technology to drive the application; we simply duplicate existing paradigms for learning without consideration of their appropriateness for or advantages of the new technology.

The technique that is presented for visualizing sampling variability reminds me of Rosling's innovative use of animation to help students to understand multivariate times series. Rosling took a familiar idea—flipbook animation—and, with the help of new technology, 'liberated the x-axis from the burden of time' (Rosling and Johansson, 2009). Wild and his colleagues take the familiar concept of shading, which has recently gained popularity via transparent colours and overplotting, and apply it to the concept of sampling variation. The blurred images that they create have great potential as a metaphor for sampling variability. As they explain, it enables students to keep their eyes on the graphical representations as they gain insight into statistical inference.

A key step is for students to move from understanding how blurry overplotted boxplots demonstrate sampling variability (Fig. 7) to successfully comparing two boxplots (Fig. 8), each representing a sample,

to detect differences between corresponding populations. Wild and his colleagues cleverly avoid several pitfalls by comparing boxplots, rather than, say, histograms. Yet, I am curious whether students have difficulty in making this transition.

Currently, I am experimenting with R (R Development Core Team, 2010) in an introductory, university level course and also am pairing physical and technological demonstrations of concepts. The students use small computational steps that parallel a physical model to construct understanding of a concept. Although I think that this approach is more effective than one function or applet that 'does it all', this claim needs to be further examined. More, it has been jarring for students to go between the computational approach and traditional textbook, and I expect that we can only have success with an integrated curriculum.

Technology is constantly changing, offering tremendous opportunities, yet threatening to make newly developed applications obsolete. In contrast, physical models seem timeless. We may ask whether the benefit of uncovering new paradigms for learning offset the burden of keeping abreast of these continual changes. The excellent contributions of Wild and his colleagues demonstrate that with careful design the benefits clearly outweigh the disadvantages.

Sastry Pantula and Ron Wasserstein (*American Statistical Association, Alexandria*)

Just as the nurturing of mathematical thinking begins early, so must statistical reasoning. It is imperative that statistics is included in a 21st-century education for all students, but many K-12 teachers have not received the necessary training at the appropriate level to teach statistical literacy in their classrooms. Sound statistical reasoning skills take time for students to develop and cannot be honed in a single course. Foundational statistical concepts should be introduced and nurtured in the elementary grades and strengthened and expanded throughout the middle school, high school and post-secondary grades. We applaud the discussion of Wild, Pfannkuch, Regan and Horton regarding the necessity of statistics being appropriately taught at the school level and the need for statisticians to care about and be involved in the statistical education of our school children.

As the authors discuss, characteristics of effective statistics education including data analysis and probability are described in Franklin *et al.* (2007) (`www.amstat.org/education/gaise`). These include teaching statistical problem solving as an investigative process, promoting statistical thinking through hands-on problem solving and concept activities, emphasizing concepts, using real world data, stressing the importance of context, recognizing variability and using appropriate technology to emphasize concepts and to support learning. Students must be taught to think statistically, in part by learning to ask key questions about scientific investigations to understand their reported conclusions and to offer an informed opinion about their legitimacy. Franklin *et al.* (2007) provide a framework to describe what is meant by a statistically literate high school graduate and provide steps to achieve this goal of a statistically literate citizenry. We appreciate the authors' discussion of how to teach the basic tenets of statistical inference at the K-12 level to guide students further to develop statistical reasoning over time.

The American Statistical Association recognizes the need for high quality teacher preparation and professional development in statistics. Well-trained teachers should have tools and training to teach statistical concepts effectively. To enhance teacher training and student learning, teachers and teacher educators need access to materials supported by statistics education research. Statisticians and statistical societies need to work to improve statistical instruction and literacy through policy and outreach efforts with education departments, teacher preparation and professional development programmes. As we enhance the statistical reasoning of teachers and students, we shall take an important step towards enabling our children to thrive in our data-driven society.

Emanuel Parzen (*Texas A&M University, College Station*)

I strongly support the authors' goals for statistical education.

(a) To motivate students about the practical value of learning statistical reasoning teach that success in life (all aspects) is achieved through understanding (rather than memorizing) formulae that are useful to answer practical questions. Learn methods 'just in time' to solve a practical problem, posed in words. Require students to formulate and solve statistical problems ultimately to answer original practical questions. Explain that each step is obtained by reasoning (applying universal strategies), not memorization.

(b) Frontiers of methods and applications can be taught if what we teach extends to more advanced problems. The competitive advantage of statisticians is that they extend to complex data methods that work for simple data; vice versa should be an aim of statistical education.

(c) Visual (produced by software) teaching and practice of statistical methods are best provided by scatter diagrams, slopes of lines and correlation coefficients. The two-sample problem (of data X_i size m, Y_j size n) should be introduced visually as a scatter diagram of combined data $(0, X_i), (1, Y_j)$. The least squares line joins $(0, \bar{X})$, $(1, \bar{Y})$ with slope $d = \bar{Y} - \bar{X}$; the significant difference of d from zero is judged by $T = \sqrt{\{(m+n-2)R\}}/\sqrt{(1-R^2)}$, where R is the sample correlation coefficient of the scatter diagram (calculated by using n rather than $n-1$ in sample variance). To compute confidence intervals calculate $\mathrm{SE}(d) = d/T$; plot quantiles $Q(P)$, $0 < P < 1$, of T and Z normal.

(d) The above regression dummy variable approach to the two-sample problem helps to introduce regression for continuous X and Y; logistic regression.

(e) An alternative to the dot plot (of sample X_i size n) is to plot the sample quantile function (it works for any sample size; 0–1 data) defined as scatter plot $((i-0.5)/n, X(i;n))$; $X(i;n)$ are ordered values (add boxplots and fences; interpret tails and symmetry). The rotating plot is the sample distribution function.

(f) All statistical education, including introductory education, should learn to solve problems by 'analogies between analogies' (links between statistical problems in different fields and different applications, and links between different parts of the course).

(g) A unifying idea of frontier statistical inference methods is to calculate the optimal function that minimizes a (information) distance from a specified function, subject to inner product constraints that provide parametric representation, and estimating equations for parameters of the optimal function. For the model selection problem, choose constraints that best fit data.

Brian Pink (*Australian Bureau of Statistics*)

H. G. Wells wrote,

'statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write'.

The authors have rightly placed the primary objective of statistical education in schools to be the learning of life skills so that students may reason and understand the use of statistical evidence, regardless of the career that they will pursue. It is indeed important that students understand basic design issues and may question issues relating to the relevance and quality of data, presentation and practical importance of the results, etc. The current curriculum tends to restrict the use of statistics to applying tools for summarizing data and making conclusions. The focus of the paper is on inferential thinking but readers must not be distracted from the importance of the wider context of sound statistical reasoning. Indeed addressing the questions 'is the data set fit for the purpose of the investigation?' and 'are the conclusions justified?' should be an important part of the journey.

The approach to teaching statistical inference that is advocated in the paper would rely on students being able to look at graphs and to interpret information. Such skills are not always found in students in early years of high school, and it would be useful if the authors could clarify how thinking relating to the data as visualized in graphs could be taught well. The authors touch lightly on the 'descriptive game' but to play the 'inference game' students need to be able to use graphical tools to describe and understand the data.

We strongly support teaching the 'minimum set of the biggest ideas of statistical inference'. Should the 'minimum set' include extending from a sample to an infinite population? We would welcome the authors' comment on how this abstract concept could be included in a manner that is underpinned by the principles proposed.

The authors made much effective use of automation in data presentation to help to draw out the message in an understandable manner. At the Australian Bureau of Statistics we have provided various interactive tools on the Web site to help to display visually the data that are disseminated (see `http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Interact+with+our+data`). Teachers should, however, encourage students to develop their own presentation tool to tell the story behind the data. Paradoxically, if our teaching of statistical thinking and inference is successful, we would not overly rely on the availability of presentation tools!

I thank the authors for a thought-provoking paper.

J.-F. Plante (*HEC Montréal*) **and N. Reid** (*University of Toronto*)

We congratulate the authors on a very original and dedicated programme of improving statistical teaching in schools. As the authors note, this may help us to recruit more, and stronger, students to our discipline. Just as important, it is part of our social role to improve statistical literacy in the population.

Too often, as educators and commentators, we describe and illustrate a list of errors in inference, a list of 'don't dos'. Although we may be correct, we rarely follow this critique by a similarly long list of acceptable approaches. Too often, the final answer is 'The solution is hard and you may learn it some day, if you choose the right optional course'. We see very positively the proposal of the authors to go ahead with simple forms of inference, leaving discussions about some important issues (and their solutions) to a later time. As students try different problems, they will gain confidence in their capacity to analyse data, and they will discover for themselves the need for more suitable models.

At the University of Toronto, we taught a first-year reading course on statistics to students with little background in statistics or mathematics. The outline of the course was intentionally vague, and stories in the news were used as pretexts to discussions of statistical methodology. We found that linking statistical methods to current stories improved the interest and the understanding of the students. We also found that explanations of puzzling or new concepts always drew more attention when discussed in the context of a question from a member of the class. The mathematical content of our course was minimal, but experience makes us believe that it is possible to gain a very intuitive understanding of traditional statistical concepts such as variability, confidence intervals, multiple comparisons and statistical significance.

Most readers of this paper will be trained statisticians who understand the subtleties of the discipline. One challenge that is not discussed is the feasibility of implementation at a larger scale, involving teachers with different levels of ability in statistics. A natural next step may be to think of optimal ways to train teachers to ensure that they are comfortable in the delivery of the material and aware of important subtleties. This should be achievable—our elementary music teachers were not virtuosos, but they did a great job at teaching us the basics—but it deserves careful planning.

Donald B. Rubin (*Harvard University, Cambridge*)

The target article displays substantial wisdom when criticizing current methods for teaching and conveying basic ideas of statistical inference. A rather dramatic example of the confusion that can be created by traditional frequentist 'thinking under the null' is a current US legal case in which the defendant was convicted of a federal crime, basically for not adhering to a strict interpretation of significance tests, when he provided a press release concerning the possible mortality benefit of his company's product for treating an inevitably fatal, swiftly advancing disease (United States *versus* Harkonen, 2010 U.S. Dist. LEXIS 75528 (N.D. Cal. July 27th, 2010)).

To avoid such confusion, however, my approach is distinctly different from that proposed in the paper, although perhaps complementary. We should start the inferential process with clear statements of scientific objectives and associated estimands, and a precise description of the hypothetical data set from which we would simply calculate the estimands. Such a data set could be composed mostly of unobserved or even unobservable values, but it defines the objects of inference. Then we should think about the mechanism that led to observed values being observed and unobserved values being missing (as in Rubin (1976)). To me, statistical inference concerns estimating missing values from observed values rather than pondering p-values, and one cannot even begin to do this without assuming something about the process that created the missing values: was it, for example, simple random sampling or was it a censoring mechanism?

This inferential process is most directly conceptualized as finding the probability distribution of missing values given observed values and scientific assumptions, i.e., formally, finding the posterior predictive distribution of the missing values, whence the posterior distribution of the estimands can be calculated. There is no need to teach technical details of this Bayesian approach; instead we can use ideas based on simulating the unknowns under scientifically motivated models. We should not avoid discussing the scientific context of any statistical problem, because this can greatly affect the resulting inference.

The approach of distributionally filling in (or multiply imputing (Rubin, 2004)) the missing values is intuitive and reveals the natural uncertainty of inference. The standard frequentist approach, which treats some known functions of the observed values as unknown (e.g. sufficient statistics) and treats some unknowns (e.g. parameters) as known, leads to the confusing double-negative logic of 'failing to reject the null hypothesis' etc.

After understanding this natural approach to access uncertainty of inference, then we can teach the importance of evaluating operating characteristics of such procedures. We should rarely, if ever, fully trust the veracity of the models used to impute the missing values, and frequency-based evaluations of them can be extremely helpful, as argued in Rubin (1984).

Richard L. Scheaffer (*University of Florida, Gainesville*)

I congratulate the authors on their original thinking on key issues surrounding statistics in schools; the

emphasis on data exploration must be expanded to include basic concepts of inference and professional statisticians must play a major role in establishing the primacy of statistics in schools while promoting creative thinking in its teaching.

The Quantitative Literacy Project in the USA (1980s), borrowing unabashedly from the UK's Schools Project, succeeded in raising interest in statistics through the new and exciting emphasis on exploratory data analysis. Our National Council of Teachers of Mathematics supported this effort and promoted statistics (data analysis) as part of the school curriculum. Emphasis on teaching statistics began to wane after 2000 when schools came under the pressure of testing as *the* measure of 'progress'. Statistical thinking, which is difficult to test and not yet in the mainstream, became neglected or replaced by the 'meanmedianmode' framework. Franklin *et al.* (2007) were instrumental in helping some states (where education decisions lie) to revitalize their curriculum in positive ways, and the new common core state standards (`http://corestandards.org/`) should expand the interest in statistics.

The common core standards contain a potentially strong statistics strand throughout the middle and secondary years, but the potential will be realized only if creative thinking of the type that is espoused in this paper becomes widespread. The plan of a logical progression of as few new ideas as are needed for basic inference, coupled with the use of appropriate technology while keeping an eye on the data, seems sound. Related work on pedagogy and assessment is also essential to the enterprise. Statistics should be taught as an investigative process; this approach allows that—even requires it.

I am happy to see the authors not projecting all inference towards the Procrustean t-test, as less restrictive mechanisms for inference are now accessible through technology. However avoidance of 'thinking under the null' is not entirely positive, as it may be quite natural for some to think 'would the observed difference in these medians be so large if the populations were the same?'.

The quotes from John Tukey are great; we should also keep in mind that his influence in making data analysis exciting and acceptable among statisticians allowed real statistics education to enter the schoolhouse door. Now, the statistical profession must take interest in 'school stuff' to keep that door open. Statistical thinking is too important to be relegated to at most one college course!

Milo Schield (*Augsburg College, Minneapolis*)

At the 1995 meeting of the American Statistical Association in Orlando, Robert Hogg talked extemporaneously about problems in teaching the introductory statistics course. I can still hear his voice saying, none too quietly,

> 'I'm tired of people talking about problems in the introductory course. I know there are problems. I've written about some of the problems. What I want is for someone to come up with a solution: a comprehensive solution, not just a small change. I want them to write up their solution in detail so I can see what to do differently. I want to be able to try it for myself. Then I will know how good it is.'

Bob Hogg was very perceptive. It is much easier to be a critic than a creator.

The Wild–Pfannkuch–Regan–Horton paper is all but certain to satisfy Bob Hogg's requirements. This paper does not just dwell on the problems; it advocates solutions; solutions that involve greater focus on ideas and concepts, less focus on the numerical details. And the solutions are detailed—not just broad generalizations; they form an integrated whole—not just a narrow change. Although some elements of this paper have been presented and discussed at various conferences (Statistical Reasoning, Thinking and Literacy, the International Conference on Teaching Statistics, the International Statistical Institute, American Statistical Association, etc.) by the co-authors and others, this paper integrates these ideas into a unified plan based on visual evidence.

The metaphors presenting these ideas are vivid and compelling:

> 'our scene setting metaphor for statistical inference starts with the idea that looking at the world by using data is like looking through a window with ripples in the glass';
>
> 'we limit attention to distortions that are produced by the act of sampling and sampling variation'.

The central issue and the author's key claim are presented in a straightforward way:

> 'some statisticians may be uncomfortable about using the non-overlap of individual uncertainty intervals to indicate [statistical] significance';
>
> 'we believe that this is a finer-distinction, later-refinement issue rather than a fundamental issue';

'each [of our guidelines] is statistically valid but with a limited range of applicability'.

In conclusion, if the goal is to introduce the idea of statistical inference to statistical beginners without the use of computational aids and with minimal mathematics, then arguably this paper represents the single biggest advance in several decades of concerted effort in statistical education.

Michael Stuart (*Trinity College Dublin*)

While admiring the efforts of the authors to simplify the introduction of key ideas of statistical inference, I am puzzled by their dismissal of process sampling as a base on which to build an approach to inference. The fact that repeated sampling is an integral part of process monitoring means that the key concept of the sampling distribution can be given an operational introduction through visualizing continuing process monitoring, by way of an evolving control chart, rather than the theoretical but practically implausible notion of repeated sampling of a population.

The fact that the distribution of evolving process data can be built up step by step means that a concrete view of that distribution can be built up and, with sufficient simulation, be seen to lead to a conceptual model. The fact that a stable process is the ideal in a process monitoring context and that control charts are designed to detect substantial departures therefrom means that the 'null hypothesis' concept arises naturally. Shewhart, the originator of the control chart, used 'chance causes' and 'assignable causes' of variation to denote null hypotheses and departures therefrom respectively. He used the normal model for chance causes, thus leading to his 3σ-limits.

By initially sampling one value at a time, corresponding to the so-called 'individuals control chart', and later moving to sampling several values at a time, corresponding to the $\bar{X}$-chart, the ideas of repeated sampling and standard error can be separately introduced, thus easing both tasks.

Related ideas become more accessible when introduced in this way. For example, the notion of a significance level may be identified with the control chart *false alarm rate* which has a simple operational interpretation in terms of an on-going process, with built-in repeated sampling, rather than the more obscure *probability of rejecting the null hypothesis when true*. It also facilitates emphasizing that there is more to statistical analysis than hypothesis testing, as a control chart is considerably more than a succession of individual hypothesis tests.

It is suggested that the approach which is advocated here passes all four tests listed at the end of Section 2.3 of the paper, that it lends itself to the wholly visual approach advocated in the paper and that it overcomes the reservations raised by Wild (2006), section 2.5, regarding process sampling.

More detail may be found in Stuart (2005).

Dennis Trewin (*Aranda*)

First, I congratulate the authors on an interesting and thought-provoking paper. To provide some background I was head of the Australian Bureau of Statistics from 2000 to 2006. During this period I took a strong personal interest in the teaching of statistics in schools and funded many initiatives. This was driven to a large extent by employers' concem about the supply of statisticians being considerably less than the demand. The demographics of statisticians suggested that this was going to become worse before it got better. Some employers raised their concerns at the Ministerial level, provoking the interest of the Minister and his acceptance that statistics was an important life skill.

A roundtable of key stakeholders was held to discuss the problem and it identified that the key part of the solution was with schools—students, teachers and pedagogy. Foremost was the lack of statistical training of teachers, inadequate teaching resources, uninspiring teaching methods and little knowledge of potential careers in statistics.

Although we had the interest of the Minister we had great difficulty in raising the interest of many in our profession. Many in senior academic positions did not see the problem. As a result progress was much slower than it could have been. The lesson is that there is a need for a co-operative approach from all the leaders of the statistical community. It should not be assumed. It may require some internal debate first.

I support the approach by Wild and his colleagues which is, in essence, learning by doing. Statistical theory is a real turn-off to most and not necessary to teach the statistical concepts and processes which are important in everyday life. It is the approach that is taken in our application of CensusAtSchools. This was the most successful of the Australian Bureau of Statistics's schools projects, helped by the involvement of teachers in the design of the product and lesson plans.

I believe that the focus of teacher training must be on mathematics teachers. Statistics in schools is most likely to be an identifiable part of the mathematics curriculum rather than a subject in its own right. Given

that statistics often becomes 'real' when its application can be demonstrated to students, statistical content in subjects such as geography and economics can be very useful.

To conclude, I strongly support the proposal of Wild and his colleagues that 'We need complexity reduction strategies'. This will mean moving away from conventional thinking on teaching.

The **authors** replied later, in writing, as follows.

We thank the Royal Statistical Society, Neville Davies and the Royal Statistical Society Centre for Statistical Education, John MacInnes, Peter Holmes and everyone who participated in the discussion of our paper. It is very encouraging to see contributions from such a diverse spectrum of the statistical family ranging from leaders of national statistical agencies, leaders of international statistical societies and Committee of Presidents of Statistical Societies award winners to many toilers at the coalface of statistical research, the provision of statistical services, statistical education and user communities. Given its societal importance, having such a broad coalition actively working together is absolutely vital to advance the penetration and impact of statistical education.

With so many contributions it is impossible to address them all individually so we shall concentrate on several general themes, leaving most of what we agree with, and some of what we disagree with, to pass without comment. Although there was plenty of support for what we are trying to do there was a spread of opinion over details, and particularly over scope.

Misinterpretations and the tyranny of page limits

Our paper combined a general appeal for engagement in reconceiving the broad spectrum of statistics in much more accessible ways with a detailed example. We detailed work in the area where we have so far done most of our own 'reconceiving'—conceptual pathways to classical statistical inference. Phrases such as 'We limit attention to ...', which were intended to relate to the scope of exposition in a page-limited paper, were often misinterpreted as limitations on the totality of what students can and should do. A second misinterpretation arose when we talked about ' "statistical inferences" as the territory addressed by confidence intervals, critical values, *p*-values and posterior distributions'. Here we were simply defining the meaning we were ascribing to the *phrase* 'statistical inference' (consistently with books that include it in their titles). We certainly did not mean to claim that this was the totality of extracting meaning from data.

The two misinterpretations above led to 'There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy' accompanied by input about what they were. The unexpected benefit has been wide-ranging discussion about what is important in statistics education, including many areas that are indeed more important than classical inference. Areas discussed included consideration of *biases* and other *non-sampling errors* (Holmes, Greenland, Goldstein, Louis and MacInnes), *more typical* but complex forms and aspects of *data production* including *process data* (Holmes and Stuart), *dependence structures* (Goldstein) and *missing data* (Rubin), *experimental versus observational data* and *causation* (Hewson), approaches based explicitly on probability including *Bayesian* (Aitkin, Hewson and Lindley), imperatives for *official statistics* (Camden), *power* and *levels of confidence* (Holmes) and the irrelevance of classical inference for very large *n* data sets (Nicholson). There were also some questions and differences of opinion about pedagogy. We shall pick up as many of these issues as space permits and conclude our response with themes raised about agencies, statistical organizations and school curricula in the advancement of statistical literacy.

Just one part of a much wider curriculum

Our work on the inference problem addresses a detailed part of a much wider statistics curriculum in New Zealand (NZ) in which statistics is taught in every year of schooling from age 5 to about 18 years as part of the curriculum area that is now called 'Mathematics and Statistics'. This statistics curriculum (`www.censusatschool.org.nz/2008/documents/new-nz-stats-curriculum.pdf`) is almost certainly the most comprehensive in the world and covers most of the above, whereas those areas that are not addressed are probably much too advanced to be broached before university. Why did we work and report on statistical inference?: because it corresponded to an important hole in our broader fabric—an important area where a good way forward had not already been worked out either by us or others.

The NZ statistics curriculum consists of three overlapping strands that go through all 13 years of schooling. These strands are *statistical investigation* (experiences in, and skills and concepts for, conducting investigations), *statistical literacy* (experiences in, and skills and concepts for, critiquing accounts of investigations by others) and *probability* (rather more traditional and building the mathematical linkages). Everything is based around the *problem–plan–data–analysis–conclusions* (PPDAC) cycle; see MacKay and

Oldford (2000) and Wild and Pfannkuch (1999)). Over the course of their schooling, our students will have gone around the PPDAC cycle hundreds of times, using it both as a guide to planning and conducting investigations, and also as a framework for interpreting and critiquing the work of others. (What is the problem that they were trying to address? What were their questions? How well do these questions address the base problem? What were their measures? Do they address the questions? How did they collect their data? ...). We owe this basic structure to Jock MacKay and Wayne Oldford's second-year course at the University of Waterloo in the early 1990s. It arose from their work on structured approaches to problem solving for managers in industrial process improvement. Thus NZ students are being educated to think critically about the various types of non-sampling errors and resulting biases when they think their way through the 'plan' and 'data' steps of the cycle; likewise for issues of measurement. This contributes to a greater or lesser degree of trust in anything the data might say. They think about more technical matters (including inference) under 'analysis'. The key issue of confounding is considered at the interface between 'analysis' and 'conclusions'. A collection of process diagrams and other representations of aspects of statistical thinking can be found at `www.stat.auckland.ac.nz/~wild/StatThink`.

Complexity and frameworks

What these discussions have begun to highlight is the complexity and subtlety of statistical thinking as it interacts with the real world. Our deep-seated desire to build all of this at once collides violently with the realities of small working memory and cognitive overload that cognitive psychologists warn about and all students (but only perceptive teachers) experience every day. And yet somehow we must try to build a very rich fabric of interlinked concepts in a way that the right nodes will be activated at the appropriate times. Our way forward relies on clustering concepts into smaller, more manageable sets that share fairly well defined spheres of influence and then finding strategies (e.g. PPDAC) to build bridges between these clusters which facilitate the right clusters being activated at the right time. Several things are clear: by attempting to do everything at once we shall achieve nothing; cognitive overload is seldom more than a hair's breadth away; and even being entirely correct is unhelpful when bought at the expense of being comprehensible.

Students in NZ are wrestling with issues of design and bias from early on in their studies of statistical literacy. As a result, when we start to introduce issues of informal inference in a stripped-down manner, they have a framework to integrate this with the larger questions that are raised by several of the discussants. As they proceed through the staged development that we have laid out, they continue to sharpen their questions and ensure that they do not get a precise answer to the wrong question.

Competing imperatives

Ours is a world of competing absolute imperatives with no 'silver bullets': only well balanced trade-offs. All our swords are double edged. Even the now accepted belief in using interesting real data has its negatives. On the one hand it is absolutely necessary for motivation, for making it obvious that certain things really are worth thinking and learning about. Additionally most of the PPDAC cycle simply cannot be experienced without real data with rich context (Wild, 1994). But on the other hand our recent research and experience say that, when trying to reason from data, students and their teachers do not know when to pay attention to *context* and when to pay attention to *pattern*. Technical statistics, including classical inference, is concerned with reasoning from pattern. Meaning, explanations of patterns thought likely to persist (Holmes) and much of critique draw most heavily on context. Statisticians rapidly shuttle between these two spheres (Wild and Pfannkuch, 1999; Cobb, 2007), something extremely subtle with a long encultural time. Consequentially, when trying to build the concepts of statistical inference by using data with a context that is too rich and compelling, teachers and students tend to go off on tangents into the context sphere at the very point where they are trying to form linkages in the pattern sphere. Thus these linkages may neither be coherently presented nor received. Timely and appropriate activation of context *versus* pattern is a critical area for researchers to develop new scaffolds.

Democratizing statistics

It was gratifying to have the reading of our paper used in the launch of the Society's 10-year statistical literacy campaign 'getstats'. As with such campaigns, the motivation for our own work is the democratization of statistics so that the benefits of statistical conceptions and modes of thought can be brought to the widest possible segments of society (Cook, Pink, Pullinger and Camden). But, since the 'traditional road to statistical knowledge is blocked, for most, by a formidable wall of mathematics' (Efron and Tibshirani as quoted by Fisher), we are seeking non-mathematical, maximally accessible, ways of conveying ideas. This precludes starting inference from probability (Aitkin and Louis) and going immediately to Bayesian conceptions (Lindley, Hewson, and Stander and Moyeed; we do look forward to their work on Bayesian

visualization). Models of dependence structures (Goldstein) and the subtle issues that were discussed by Rubin raise even higher barriers to mass entry, whereas process data (Stuart), with its lurking menaces of dependence and non-stationarity, are inherently more complex than sampling or randomization data. Ideas about these more complex matters should be easier to build if some solid intuitions have already been established in simpler, if less frequent (Holmes), contexts. Meng's metaphors of learning swimming in swimming pools but glimpsing oceans, etc., are gloriously apposite.

Description and inference

Although we agree with Gelman and Maclnnes that in reality description and inference are intimately interlinked, developmentally we have to pull these things apart (Pfannkuch *et al.*, 2010). In part this is because first people need 'to learn to see' and this does not come quickly even with what statisticians consider trivially simple displays. A focus on description helps with 'learning to see' and the distinction also helps to cement 'what I'm seeing differs somewhat from the underlying reality'. We use the triggers (Fig. 1 in Pfannkuch *et al.* (2010)), '*I notice* . . . ' (to trigger thoughts about what I am seeing), '*I wonder* . . . ' (to trigger thoughts about what of this might reflect aspects of the deeper reality), '*I worry* ...' (about data relevance, quality and alternative explanations) and '*I expect* . . . ' (what I expected to see and what surprises me). We shall appropriate Holmes's excellent 'What might have caused these data to be as they are?'.

Key concepts of inference

We agree with Agresti that confidence intervals are more generally useful than tests. The most critical messages are, in decreasing importance for reach, and beginning by adapting Box on models,

(a) all estimates are wrong but some are sufficiently close to be useful (similarly measures),
(b) every estimate should be accompanied by a measure of uncertainty,
(c) ideas about what uncertainties these measures account for,
(d) beginning ideas about how to obtain them and
(e) technical understandings and competence.

We would like everyone to understand the first and to feel aggrieved when not provided with the second, and many to be steeped in what questions to consider when addressing the third. For the fourth we favour the bootstrap, with facilitates handling measurement data and category data (Agresti and MacGillivray) simultaneously. But it is all a question of how fast you can get there and that seems to be highly dependent on maturity. For the ages of students whom our teachers were working with we found that we had to slow the process from sampling variation to confidence intervals down and first to try to make a call on direction. For many young students, self-discovery that 'the data can get it backwards' is revelatory, as is the prior self-discovery that 'I can get useful information about a population by taking a sample' (Louis). And as part of these discovery processes our students do learn about being wrong (Holmes) and acting in ways that prevent this happening too frequently. These classroom implementation issues (Borovcnik, King and Woodford) and assessment (Kapadia) will be reported on separately. Since we wrote this paper Chris Wild has been working on dynamic visualizations for the simple bootstrap with exciting results. The direct visual links between the effects of sampling as a dynamic visual process and mimicked by the effects of bootstrap resampling as a dynamic visual process make bootstrap confidence intervals quite compelling; see `www.stat.auckland.ac.nz/~wild/BootAnim`. 'Do they work?' Simulate and see!

We shall teach 'thinking under the null' and significance (Bowman and Scheaffer, and despite Bayesian and other objections of Hewson, Lindley and Rubin) in the last year of school, but in modules on data from designed experiments (where the idea of an intervention making absolutely no difference is plausible) using randomization tests so that the analysis involves no mathematical machinery and is directly linked to the data production mechanism.

As to the many challenges that have been issued (Ainley and Pratt, Borovcnik, Garfield and Zieffler, Kapadia and Nolan), brevity precludes saying more but many correspond to work that is already in train or on the drawing board and we encourage others to join us in trying to answer these very important questions.

Curricula and the advancement of statistical literacy

The desire for significant whole-society increases in statistical literacy is strongest in those currently or formerly in leadership positions in statistical agencies (Pink, Trewin, Cook and Pullinger) and to a significant if lesser extent in those who lead or have led Societies (e.g. Pantula and Wasserstein, Scheaffer, Fisher, Louis and MacGillivray). This is not surprising as these are all roles which demand big picture vision, far beyond 'the sorts of statistics I do and the sorts of students I see'. As National Statisticians Trewin (Australia) and Pink (NZ and then Australia) have really stood out for their personal investments in pushing education agendas to achieve such ends. Pink's successor heading Statistics New Zealand, Geoff

Bascand, has continued this tradition and 20 years of lobbying of educational officials by Mike Camden has paid real dividends.

We believe that the only feasible road to substantial increases in whole-society statistical literacy is an appreciable presence of statistics in compulsory school curricula wisely used; anything less amounts to tinkering around the margins. It might be valuable tinkering but, in terms of percentage reach, it is tinkering nonetheless.

Concern has been raised about the training and ability of teachers to teach statistics and the need for development (Bowman, Plante and Reid, Trewin, Fisher, Gould, Pantula and Wasserstein, and Meng). Unfortunately, substantial investment in any of these things follows presence in the actual *assessed* curriculum; it does not precede it. If we wait until everything is perfect in terms of teacher preparation nothing will ever happen. We must either trust that intelligent people, provided with appropriate support coupled with research evidence (Borovcnik and Kapadia), can learn to do these things or conclude that widespread statistical literacy is a just nice dream that can never happen.

Although some things can be learned from the NZ experience its small size makes it too special. Australia has the large-society complications of states with the power to make their own rules. The following are gleaned from experiences in Australia (Trewin and Fisher), NZ and even some in the UK. The processes of curricular reform are almost entirely political. Doors of opportunity momentarily open then rapidly close for years, perhaps decades. The national statistical society and the national statistical offices need to be allies who speak with one voice with commitment from the top of both organizations. Each organization has legitimacy with different stakeholders and can open different doors. Although connections at a political level are very useful they are also very transient. Their most important advantage may be in receiving a sympathetic hearing from key officials who then must be convinced that statistics is critical to societal advancement, is different from mathematics with almost orthogonal imperatives and is *the province of an entirely different community* (see Fisher). This is an on-going challenge as, even in the bureaucracy, people and roles change. Finally it is critical to have ears to the ground that will detect the earliest rumblings of a curriculum movement. As soon as the first tremors are felt, start running fast and pray that you have not already been shut out by all the competing voices.

References in the discussion

Advisory Council on Mathematics Education (2010) *Post-16 in 2016*. London: Royal Society.

Bolstad, B. (2007) *Introduction to Bayesian Statistics*. Hoboken: Wiley.

Borovcnik, M. and Kapadia, R. (eds) (2009) Special issue on "Research and developments in probability education". *Int. Electron. J. Math. Educ.*, **4**, no. 3.

Bowman, A. W., Crawford, E., Alexander, G. and Bowman, R. W. (2007) `rpanel`: simple interactive controls for `R` functions using the `tcltk` package. *J. Statist. Softwr.*, **17**, no. 9.

Brillinger, D. R. (2002) John W. Tukey: his life and professional contributions. *Ann. Statist.*, **30**, 1535–1575.

Bruner, J. (1996) *The Culture of Education*. Cambridge: Harvard University Press.

Chambers, J. (1993) Greater or lesser statistics: a choice for future research. *Statist. Comput.*, **3**, 182–184.

Cobb, G. W. (2007) The introductory statistics course: a Ptolemaic curriculum? *Technol. Innovns Statist. Educ.*, **1**, 1–15. (Available from `http://escholarship.org/uc/uclastat_cts_tise`.)

Connor, D., Davies, N. and Holmes, P. (2006) Using real data and technology to develop statistical thinking. In *Thinking and Reasoning with Data and Chance* (eds G. Burrill and P. C. Elliott). National Council of Teachers of Mathematics.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.

Fisher, N. I. (2010) Statistics education in Australia. *SSAI Newslett.*, Mar. (Available from `http://www.statsoc.org.au/objectlibrary/547?filename=SSAI-Newsletter-Mar2010-Final-Webready.pdf`.)

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M. and Scheaffer, R. (2007) *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: a Pre-k-12 Curriculum Framework*. Alexandria: American Statistical Association.

Geneletti, S., Richardson, S. and Best, N. (2009) Adjusting for selection bias in retrospective case–control studies. *Biostatistics*, **10**, 17–31.

Grandin, T. (2006) *Thinking in Pictures*, 2nd edn. London: Bloomsbury.

Greenland, S. (2005) Multiple-bias modelling for analysis of observational data (with discussion). *J. R. Statist. Soc.* A, **168**, 267–306.

Greenland, S. (2009) Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statist. Sci.*, **24**, 195–210.

Gustafson, P., Le, N. D. and Saskin, R. (2001) Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*, **57**, 598–609.

Hattie, J. (2009) *Visible Learning*. Abingdon: Routledge.
Hodges, J. L., Krech, D. and Crutchfield, R. S. (1975) *StatLab: an Empirical Introduction to Statistics*. New York: McGraw-Hill.
Holmes, P. (1997) Assessing project work by external examiners. In *The Assessment Challenge in Statistics Education* (eds I. Gal and J. Garfield), pp. 153–164. Amsterdam: IOS Press.
Hotelling, H. (1948) The teaching of statistics. *Ann. Math. Statist.*, **19**, 95–115.
Kahneman, D., Slovic, P. and Tversky, A. (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kapadia, R. and Borovcnik, M. (1991) *Chance Encounters: Probability in Education*. Dordrecht: Kluwer.
King, T. (2011) Statistics in Society: three case studies in the UK. *Applications and Policy Working Paper A11/01*. Southampton Statistical Sciences Research Institute, University of Southampton, Southampton. (Available from `http://www.soton.ac.uk/s3ri/publications/details.php?id=170`.)
Kumar, K. (2010) How to make teaching of statistics more effective in business schools. In *Proc. International Academy of Business and Economics Summer Conf., Bangkok*, vol. 7, pp. 60–64.
Lash, T. L., Fox, M. and Fink, A. K. (2009) *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer.
MacKay, R. J. and Oldford, R. W. (2000) Scientific method, statistical method and the speed of light. *Statist. Sci.*, **15**, 254–278.
Mandel, M. and Rinott, Y. (2009) A selection bias conflict and Frequentist versus Bayesian viewpoints. *Am. Statistn*, **63**, 211–217.
Marriott, J. M., Davies, N. and Gibson, E. (2009) Teaching learning and assessing statistical problem solving. *J. Statist. Educ.*, **17**, no. 1. (Available from `http://www.amstat.org/publications/jse/v17n1/marriott.html`.)
Meng, X.-L. (2009a) Statistics: your chance for happiness (or misery). *Harv. Undergrad. Res. J.*, **2**, 21–26.
Meng, X.-L. (2009b) AP statistics: passion, paradox, and pressure (part I). *Amstat News, Dec.*, 7–10.
Meng, X.-L. (2009c) Desired and feared—what do we do now and in the next 50 years? *Am. Statistn*, **63**, 202–210.
Meng, X.-L. (2010a) AP statistics: passion, paradox, and pressure (part II). *Amstat News, Jan.*, 5–9.
Meng, X.-L. (2010b) Rejoinder: Better training, deeper thinking, and more policing. *Am. Statistn*, **64**, 26–29.
Molitor, N.-T., Best, N., Jackson, C. and Richardson, S. (2008) Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: application to low birth weight and water disinfection by-products. *J. R. Statist. Soc.* A, **172**, 615–637.
Petocz, P. and Reid, A. (2010) On Becoming a Statistician—a qualitative view. *Int. Statist. Rev.*, **78**, 271–286.
Pfannkuch, M., Regan, M., Wild, C. and Horton, N. J. (2010) Telling data stories: essential dialogues for comparative reasoning. *J. Statist. Educ.*, **18**, no. 1. (Available from `http://www.amstat.org.publications/jse/v18n1/pfannkuch.pdf`.)
R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rosling, H. and Johansson, C. (2009) Gapminder: liberating the x-axis from the burden of time. *Statist. Comput. Graph. Newslett.*, **20**, 4–7.
Roth, W.-M. (2004) Emergence of graphing practices in scientific research. *J. Cogn. Cult.*, **4**, 595–627.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.
Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys*, appendices 1 and 2. New York: Wiley.
Schools Council Project on Statistical Education (1980) *Statistics in Your World*. London: Foulsham.
Smith, A. F. M. (2004) *Making Mathematics Count*. London: Stationery Office.
Stuart, M. (2005) Mathematical thinking versus statistical thinking; redressing the balance in statistical teaching. *Technical Report 05/07*. Department of Statistics, Trinity College Dublin, Dublin. (Available from `http://www.scss.tcd.ie/disciplines/statistics/tech-reports/0507.pdf`.)
Tukey, J. W. (1953) The growth of experimental design in a research laboratory. In *Research Operations in Industry*, pp. 303–313. New York: King's Crown Press.
Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S. and Thompson, S. G. (2009) Bias modelling in evidence synthesis. *J. R. Statist. Soc.* A, **172**, 21–47.
Utts, J. M. and Heckard, R. F. (2007) *Mind on Statistics*, 3rd edn. Brooks–Cole.
Vansteelandt, S., Goetghebeur, E., Kenward, M. G. and Molenberghs, G. (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sin.*, **16**, 953–980.
Wild, C. J. (1994) On embracing the 'wider view' of statistics. *Am. Statistn*, **48**, 163–171.
Wild, C. J. (2006) The concept of distribution. *Statist. Educ. Res. J.*, **5**, no. 2, 10–26.
Wild, C. J. and Pfannkuch, M. (1999) Statistical thinking in empirical enquiry (with discussion). *Int. Statist. Rev.*, **67**, 223–265.
Zieffler, A., Garfield, J., delMas, R. and Reading, C. (2008) A framework to support research on informal inferential reasoning. *Statist. Educ. Res. J.*, **7**, no. 2, 40–58. (Available from `http://www.stat.auckland.ac.nz/serj`.)