COMPARING BOX PLOT DISTRIBUTIONS: A TEACHER'S REASONING

MAXINE PFANNKUCH
The University of Auckland, New Zealand
m.pfannkuch@auckland.ac.nz

ABSTRACT

Drawing conclusions from the comparison of datasets using informal statistical inference is a challenging task since the nature and type of reasoning expected is not fully understood. In this paper a secondary teacher's reasoning from the comparison of box plot distributions during the teaching of a Year 11 (15-year-old) class is analyzed. From the analysis a model incorporating ten distinguishable elements is established to describe her reasoning. The model highlights that reasoning in the sampling and referent elements is ill formed. The methods of instruction, and the difficulties and richness of verbalizing from the comparison of box plot distributions are discussed. Implications for research and educational practice are drawn.

Keywords: *Statistics education research; Box plots; Distributional reasoning; Secondary statistics teaching; Informal statistical inference*

1. OVERVIEW

Traditionally, statistics instruction focuses on the construction of graphs, which results in students not knowing why graphs are constructed in the first place (Friel, Curcio, & Bright, 2001). Graphs are frequently used as illustrations of data rather than as reasoning tools to learn something new in the context sphere, gain new information, or learn from the data (Wild & Pfannkuch, 1999; Konold & Pollatsek, 2002). A shifting of the instructional focus to reasoning from distributions for the purposes of making sense of data, for detecting and discovering patterns, and for unlocking the stories in the data, presents many challenges. In particular, a challenge is to understand the nature and type of reasoning involved when making informal inferences from sample distributions about population distributions. Without research that attends to the complexity of informal inference and its role in the building of concepts towards formal statistical inference, statistical inferential reasoning may continue to elude many teachers and students. There is a need to understand inferential reasoning about many different types of distributions but this paper will focus on the comparison of box plot distributions. Box plots condense, summarize, and obscure information, incorporate statistical notions such as median and quartiles, and are conceptually demanding for students (Bakker, 2004). Therefore the aim of this paper is to achieve a greater understanding of the informal inferential reasoning necessary for comparing box plot distributions through analyzing one teacher's reasoning.

Statistics Education Research Journal, 5(2), 27-45, http://www.stat.auckland.ac.nz/serj

1.1. REVIEW OF RELEVANT LITERATURE

Exploratory data analysis (EDA) gave rise to a number of new graphical techniques. Tukey (1977) invented box plots as a powerful way of summarizing distributions of data to allow visual comparisons of centers and spread through the five-number summary (minimum, lower quartile, median, upper quartile, maximum), which divides the data into four equally sized sections. Further refinements can be made to basic box plots by visually representing extreme values or outliers, means, and significant differences. Basic box plots are introduced to students from as young as 12 in the USA to as old as 17 in France, whereas some countries, such as China and Israel, do not have them in the curriculum (Bakker, Biehler, & Konold, 2005). In New Zealand box plots have been in the curriculum for 14-year-olds for the last 20 years.

When comparing two box plot distributions traditional instruction assumes that inferences will *not* be drawn and hence focuses on describing features of box plots. Recent changes, however, to Year 11 (15-year-old) assessment in New Zealand assume that conclusions will be drawn from visual comparisons (Pfannkuch & Horring, 2005). At this year level students have not been exposed to confidence intervals and significance testing to draw conclusions; rather their reasoning must involve informal inferential reasoning. That is, in the case of box plots, being able to infer that one group is generally greater than a second group, or that no distinction can be drawn, based mainly on looking at, comparing, and reasoning from box plot distributions. The question arises as to what elements of reasoning are necessary to draw informal inferences.

Because formal inferential reasoning focuses on the centers of distributions the question arises as to how to scaffold students' understanding towards viewing centers as being representative of a set of data. Konold and Pollatsek (2002) note that research has demonstrated that students know how to compute averages but few use averages to characterize a dataset or to make comparisons between datasets. Such a situation does not provide conceptual foundations for the development of students' inferential reasoning. Furthermore, Konold and Pollatsek (2002) identify four views of average – signal amongst the noise, data reduction, fair share, and typical value – which are dependent upon the goal the person has in mind when using an average. They argue that all goals are valid but if students do not have the "signal amongst the noise" view of average then this can result in a reluctance to use averages to compare two groups, a fact noted by Biehler (2004) in his research on box plots. Therefore, adopting the position that the middle part of the data usefully characterizes the group and that the middle parts of the distributions should be compared is a necessary element of the reasoning process.

Box plots illustrate the signal (the center) and noise (the spread of data from the center) in their representation yet according to Biehler (2004) the interpretation of spread can result in five different views, namely: location information, regional spreads and densities, global spread as a deviation from the median, median upward and downward spread, and classification information. Whatever view is taken, a spread element of reasoning must include notions of comparing variability within and between box plots. Biehler (2004) and Friel (1998) identified that the cut-off points represented in the box plot result in students using these for comparing distributions. That is, the nature of the representation leads students to argue intuitively with the data by comparing equivalent and non-equivalent five-number summary points. Thus another element of reasoning associated with comparing box plot distributions is of the summary type. Other elements of reasoning identified by Biehler (2004) as lacking in students are the "shift" interpretation and intuitions about sampling variability. He describes the "shift" element, where all the five-number summary values are higher for one box plot compared to the

other, as being an essential notion in comparison whereby students can determine the amount of the shift and the type of shift, uniform or non-uniform. If this shift type of reasoning does not work with the box plots under consideration then the comparison becomes complex (Bakker et al., 2005). Biehler's (2004) reference to sampling variability accords with Bakker (2004), who states that a key concept in developing a notion of distribution is sampling, and with Pfannkuch (2005), who believes that sampling reasoning is essential for building concepts towards formal inference. Furthermore, Bakker and Gravemeijer (2004) argue that in instruction students should experience summarizing dot plot distributions by intuitively dividing the data into groups. Such instruction can gradually develop a student habit to overlay box plots on dot plots.

Currently, studies are focused on how to introduce students to box plots and how students interpret them. There appears, however, to be no research on how teachers reason when comparing box plot distributions, nor any definitive account of how teachers or students should draw informal inferences. According to Bakker and Gravemeijer (2004) reasoning with shapes forms the basis of reasoning about distributions whereas Friel et al. (2001) refer to visual decoding, judgment, and context as three critical factors in students' abilities to derive meaning from graphs. Furthermore, Friel et al. consider that research is needed on understanding what it is about the nature of the reasoning that makes comparing datasets such a challenging task. Whatever the nature of the reasoning is, it is complex and may depend on the ability to decode representations, to attend to a multiplicity of elements represented within and between the box plots, and to make judgments.

1.2. RELATED RESEARCH

The research described in this paper is part of a larger project that is concerned with developing students' statistical thinking based on the Wild and Pfannkuch (1999) framework. In 2003, the first year of the project, informal inferential reasoning was identified as a problematic area. Focusing on the comparison of box plots, the videotape data of the classroom teaching revealed that the teacher in only one instance out of a possible eight opportunities communicated and wrote down how she would draw a conclusion from such plots (Pfannkuch & Horring, 2005). Over half the students, in an open-ended questionnaire, identified that they did not know how to draw evidence-based conclusions. An analysis of student responses to an assessment task requiring the drawing and justifying of inferences from the comparison of box plots concluded that 90% compared equivalent and 50% non-equivalent five-number summary statistics, a "summary" element; 50% mentioned the difference in the ranges, a basic "spread" element; and 30% had a very basic "shift" element of reasoning.

Realizing that drawing conclusions from the comparison of box plot distributions is not an easy task the researcher and five statisticians met in 2003 to discuss the type of reasoning that could be expected for informal inference. In teaching situations where there is no access to technology and no student experience of sampling variability, such informal inference was considered problematic not only for students but also for the statisticians (Pfannkuch, Budgett, Parsonage, & Horring, 2004). From the perspective of formal inference for the comparison of data plots the statisticians determined that there were four basic aspects to attend to in order to understand the concepts behind significance tests, confidence intervals, p-values and so forth before drawing a conclusion. These were comparisons of centers, comparing the differences in centers relative to the variability, checking the distribution of the data (normality assumptions, outliers, clusters), and the sample size effect. The discussions raised further questions as

to what types of learning experiences would develop students' inferential reasoning towards a more formal level.

Since articulating the messages contained in box plots and justifying inferences either verbally or in writing was considered difficult for both teachers and students, the idea of providing a framework to support the reasoning process was conceived. The framework would support learning in terms of what should be noticed and attended to when looking at the plots. Since Year 11 students had not been exposed to ideas of sample and population or of sampling variability and the effect of sample size, the group conjectured that perhaps students should work with clear-cut comparisons that had similar spread, no unusual patterns, and samples sizes of 30. For writing a conclusion they proposed that it should begin with the words: "These data suggest …" and that the justification should be focused on comparing the centers and on comparing the differences in centers relative to the variability. After that the students could comment on features such as variability within and between the box plots, the shapes of the distributions and compare the median of distribution *X* with the percentage of the distribution *Y* that was below it. Finally the students should check whether their conclusions made sense with what they knew from their own knowledge and consider possible alternative explanations for the findings – an explanatory element of reasoning. The statisticians and researcher also suggested dot plots should be kept with box plots (Bakker & Gravemeijer, 2004; Carr & Begg, 1994) and gave ideas on how students could experience variation (Pfannkuch, 2005).

When these ideas were presented to the teacher who was being researched, she was adamant that she wanted to deal with the inherent messiness of data where clear-cut decisions are not obvious. She also felt that some suggestions for justifying inferences and comparing features were too hard for students. At this stage, the teacher was not ready to deal with sampling variation ideas or putting dot plots and box plots together, but she was ready to try and reason from box plots. Since there seemed to be no account in textbooks and in research of how to draw informal inferences from box plots in a school teaching situation and no consensus on the statisticians' suggestions, the teacher and researcher were placed in the situation of learning in and from practice.

1.3. RESEARCH QUESTION

As part of a larger project on developing Year 11 students' statistical thinking, the following research question is addressed:

> *What reasoning does a teacher articulate when learning to communicate statistical ideas and make informal inferences from the comparison of box plots?*

2. RESEARCH METHOD

The research method is developmental in that an action-research cycle is set up whereby problematic areas are identified by a teacher and researcher through observations and critical reflections on the implementation of a teaching unit and by the researcher through analysis of student assessment responses (see Pfannkuch & Horring, 2005, for a more complete account). The teacher and researcher then discuss how the current situation might be changed for the following year when the unit is taught again.

According to Ball and Cohen (1999) actual teacher learning requires some disequilibrium since learning will only occur when existing practices are challenged. From the teacher's perspective, her practice had been challenged and hence in the 2004 implementation of the statistics-teaching unit, the teacher decided to make a conscious effort to communicate and articulate how she was looking at and what she was thinking

about when comparing two box plots. She was also aware of the need to write down the justifications for her conclusion. In this research the teacher is being put in the position of a learner in and from her practice, that is, actively learning while she is teaching. Therefore the action-research method is appropriate in such a situation.

The teacher and researcher decided before the teaching of the unit that when reasoning with box plot distributions she would refrain from using the summary element of reasoning and instead focus on the following five elements: comparison of centers, spread, the degree of overlap of the two box plots, sampling, and explanatory. The teacher decided when to introduce each element, what language she would use and how she would reason within those broad elements. After each teaching episode, the researcher and teacher had brief conversations about the type of reasoning used and what possibly to emphasize in the next lesson. Each lesson was videotaped by the researcher.

2.1. PARTICIPANT

The school in which the project is based is a multicultural, secondary girls' school. The teacher is in her mid-thirties, and has taught secondary mathematics for twelve years. In Year 10, students are introduced to the graphing of box plots. The class is taught mathematics by the teacher for four hours per week. The teacher is in charge of Year 11 mathematics and therefore, in consultation with the other Year 11 teachers, writes an outline of the content to be covered, together with suggested resources and ideas for teaching the unit. The researcher previously knew the teacher on a professional basis.

2.2. THE TEACHING EPISODES

This paper focuses on two of the three teaching episodes in which box plots were introduced and discussed. The teacher chose the tasks she gave to the students. Before the first teaching episode on comparing box plot distributions the students compared data using back-to-back stem-and-leaf plots and calculated the five-number summary for data. For homework the students were given an example from a textbook (Figure 1(a)) for which they were required to draw a back-to-back stem-and-leaf plot (Figure 1(b)) and calculate the five-number summary. The first teaching episode on constructing a box plot started at this stage. The teacher discussed and interpreted the stem-and-leaf plot, then used the five-number summary to remind the students how to draw box plots. She drew the box plot of males' pay with the class (Figure 1(c)). After the students had drawn the box plot of the females' pay she discussed the plots with them and built up a written conclusion on the board (Figure 1(d)).

In the second teaching episode the focus was on interpreting box plot distributions. The students were given a brief account of where the data had come from (Figure 2 (a)) and the dataset. They were asked to reflect on the background information and the given data and to think of questions they might pose. After the students suggested a number of questions the teacher said she had already drawn the plots for one of their questions (Figure 2(b)). The teacher articulated her reasoning from the comparison of the box plots with the class responding to and asking her questions about the data. Figure 2(c) is the conclusion she wrote on the board.

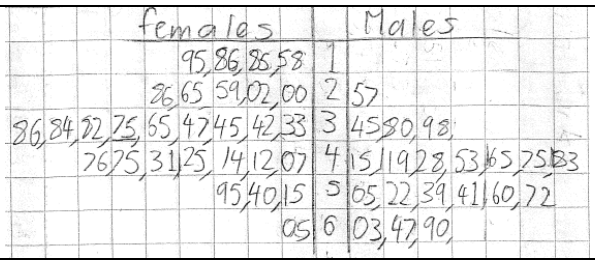

Figure 1(b)

STEM & LEAF/QUARTILES HOMEWORK

The takehome pay of 50 full time employees (30 female and 20 male) at a local firm is displayed below.

Male ($)
354 428 560 475 398 453
380 690 257 483 539 647
572 415 465 522 603 541
505 419

Female ($)
186 259 347 286 185 345
382 414 265 375 540 605
475 333 386 412 298 515
425 384 195 200 342 431
595 476 365 202 407 158

Figure 1(a) (Source unknown)

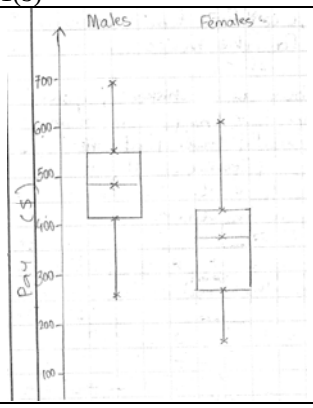

Figure 1(c)

Teacher's written conclusion was:
Each of the statistics (median, UQ, LQ) for females is lower than for the males. The female graph is clustered between the median and the UQ. The box & whisker graphs overlap but the female graph is generally lower than the males' graph. Overall, it appears that the females earn less than the males.

Figure 1(d)

Figure 1. Teaching episode one – graphs are from a student's book

Brain Size & Intelligence

A researcher collected a sample of 40 right-handed Anglo-University Students at a United States University. The subjects took subtests of the Wechsler Adult Intelligence Test. The Verbal IQ Scores, based on the test, were recorded. The researcher used Magnetic Resonance Imaging (MRI) to determine the brain size of the subjects and posed 3 questions:

Figure 2(a)

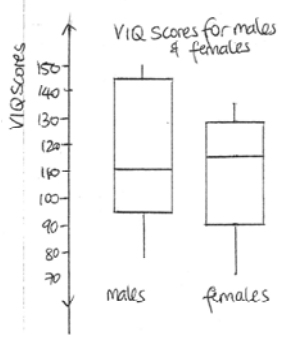

Figure 2(b): Teaching episode two

Teacher's written conclusion was:
Males' graph – slightly higher. Males' IQ is more spread than females. Interquartile range = UQ-LQ. Males IQR=39, females IQR=50. Graphs are overlapped – especially between LQ and UQ. Measure of central tendency – Female median (116) is slightly higher than males. Based on these data values, we're not certain that males have higher IQ. There is some evidence to suggest that males have higher IQ for these Uni. students.

Figure 2(c)

Figure 2. Teaching episode two – graph given to students by teacher

3. RESULTS

A qualitative analysis by the researcher of the teacher's discussion on the comparison of box plot distributions extracted ten elements of reasoning (Figure 3). Elements 4 to 7 were based on the a priori agreement, as described earlier, between the teacher and researcher on the type of reasoning elements that should be emphasized in instruction, whereas elements 1, 2, and 3 arose during teaching and the analysis. The eighth element was considered after discussion with another researcher (Tim Burgess, personal communication, 7 July 2005). It was not determined before teaching how the teacher would reason within these broad classifications. The analysis of the data suggested that the eight elements of reasoning are non-hierarchical, interdependent but distinguishable. The two moderating elements of reasoning, 9 and 10, arose from the analysis and are contained within each of the other eight elements.

ELEMENTS OF REASONING	
1. Hypothesis generation	Compares and reasons about the group trend.
2. Summary	Compares equivalent five-number summary points. Compares non-equivalent five-number summary points.
3. Shift	Compares one box plot in relation to the other box plot and refers to comparative shift.
4. Signal	Compares the overlap of the central 50% of the data.
5. Spread	Compares and refers to type of spread/densities locally and globally within and between box plots.
6. Sampling	Considers sample size, the comparison if another sample was taken, the population on which to make an inference.
7. Explanatory	Understands context of data, considers whether findings make sense, considers alternative explanations for the findings.
8. Individual case	Considers possible outliers, compares individual cases.
MODERATING ELEMENTS OF REASONING	
9. Evaluative	Evidence described, assessed on its strength, weighed up.
10. Referent	Group label, data measure, statistical measure, data attribution, data plot distribution, contextual and statistical knowledge.

Figure 3. Teacher's model of reasoning from the comparison of box plots

3.1. THE ELEMENTS OF REASONING

The goal of the teacher was to make an informal inference about populations when comparing sample distributions and to justify that inference. Since informal inferences were being drawn, visuo-analytic thinking was used by the teacher. She gradually built up, in her communication, the multifaceted ways in which she looked at and interpreted the comparison of box plots. Within some elements there are sub-elements, not all of which are illustrated below. The focus of the analysis is on her reasoning within each element as she learns more about the data under consideration. Her reasoning, however, is linked to how she teaches and therefore consideration is given to instructional methods in the analysis. It should be noted that the teacher uses the term graph when talking about stem-and-leaf plots or box plots but to be consistent the analysis of her reasoning will use the term plot.

Element 1: Hypothesis Generation Hypotheses may be formed at the beginning of an investigation before data are collected or on inspection of a given dataset, or during an investigation when analyzing a graph, or at the end of an investigation. In teaching episode one (see Figure 1) the teacher discussed the back-to-back stem-and-leaf plots before showing the students how to construct the box-and-whisker plots. In the hypothesis generation element she communicates that the inference "males earn more than females" does not capture the actual story in the data.

> Teacher: All right, what's the first thing that strikes you when you looked at this graph [Figure 1(b)]?
> Student: Males earn more.
> Teacher: Males earn more – what gave you that impression from the graph?
> Teacher: Yes, the higher amounts go down, so the mass or the bulk of the graph is lower down than the females. So that gives us the impression, that the males earn more than the females. Is that true for every single person? Did every single male earn more than every single female? No. Okay. This person here, this woman here, earns $605, she earns way more than this male here $257 …[later on] … because there's a bit of an overlap, I have to be a bit more subtle about my language and say it appears from the graph that the bulk of males, appear to earn more than the bulk of the females.

To generate a data-based hypothesis, consideration is given to variability through acknowledging that the reasoning is about the group trend and not about individual cases.

Element 2: Summary In this element the five-number summary is located on the plots and equivalent summary points are compared. For example in teaching episode one:

> Teacher: Right if we were to compare each measure like the median, the females are lower than the males. Lower quartile, females are lower than the males. Upper quartile, females are lower than the males, top – right, so each of the statistics is lower for the females than it is for the males.

The box plot representation also encourages the comparison of non-equivalent summary points such as "75% of males earn more than 75% of females," which is a comparison of a lower quartile with an upper quartile. The teacher briefly mentioned that "a quarter of the ladies, women, are earning more than these guys" in teaching episode one. Sometimes her focus is specifically on the comparison of the medians as was the case in teaching episode two:

> Teacher: The next thing that I think is a really important factor for helping me make up my mind is the measure of central tendency or average. What have I got to help me figure out what the central tendency is?
> Student: The median.
> Teacher: The median, so next I'm going to look at my median because that's the middle of the data, that gives me where the bulk of the data is and I've got females slightly higher than males. So I'm going to say the feature that I've noticed for the female's median I'm going to say what it is (116), is slightly higher than males.

The notion of the median being representative of the distribution or the signal amongst the noise is unclear in the communication but the teacher is drawing attention to its importance as a factor for making a decision under uncertainty. Mentioning the "bulk of the data" in terms of the median may be a misleading interpretation for these datasets.

Element 3: Shift In the shift element the plots are looked at as a whole and compared in terms of whether one is higher or further along than the other. This shift element could be incorporated into the summary element when equivalent summary values are compared but the way of reasoning is different in that for the shift element the box plots are looked at as a whole visually not as a straight comparison between values. From teaching episode two:

> Teacher: All right, one of the biggest factors that helps me make up my mind, … is whereabouts is the whole of the graph in relation to each other. And when I look at this, straight away I notice that the males' graph is a little bit higher up than the girls' graph. Pretty much?
> Teacher: You reckon it's a long way?
> Teacher: No, not a long way up, they're quite overlapped, they seem to be quite next to each other but one's a little bit higher than the other. So, I'll write that down. Male's graph is slightly higher.

At a more detailed level she compared the plots in terms of the shift of the majority with statements such as "the mass or the bulk of the graph is lower down." At no stage did she quantify the shift, preferring instead to use qualitative statements such as "slightly higher."

Element 4: Signal The signal element could be incorporated into the shift element but is given a separate category since reasoning about measures of center is important for formal inference. The signal element referring to the middle 50% of data, the box, may represent the starting point for informal reasoning about center. In her communication in teaching episode two the box appears to be used as a crude measure of the center and is represented as the "typical value" or the "signal" of each distribution. She is interested in how much overlap there is between the middle 50% of data so a sense of comparison of the "middle" is conveyed to students. In response to a student's query on the meaning of the term overlap she drew double-arrowed lines in the "central boxes" to demonstrate the term.

> Teacher: All right, I also might notice that the graphs are overlapped okay.
> Student: What does that mean?
> Teacher: Well that means that there's not one graph separate from the other graph, they're overlapping. So see this central box here and this central box here, remember that gives me the middle 50% of the data, they are quite overlapped, okay. Especially between, and I'll write this down, especially between the lower quartile and the upper quartile. They're very overlapped in this central part. Now in terms of which part of the graph gives me the most information or the most significant information okay, this middle box is most important, okay 'cause that's where the middle bulk is.

Possibly the teacher is laying down intuitive foundations for formal inferential reasoning where the difference in centers are compared relative to the variability. The drawn lines may also be conceived as visual foundations for confidence intervals for population medians.

Element 5: Spread The teacher drew attention to the spread element by focusing on comparing ranges and interquartile ranges visually and quantitatively. At a more detailed level she drew attention to the location of the data, that is, where and how the data are distributed. From teaching episode one:

Teacher: What does each of these sections represent? Because that's the highest, the upper quartile, median – because each part, there's four parts. See how I've broken the graph essentially into four parts.
Student: Is it the spread of money?
Teacher: It's the spread, yes, it's the spread of how much they get paid. So there are, of all the people in this study, 25 of them – percent sorry, 25 percent – a quarter right, are sitting in here. … Okay, in here is another quarter of the people, in here, another quarter, and in here another quarter. So if you imagine these women, and they were standing on this number line, and they would be –
Student: Squashed up.
Teacher: Squashed up, yes. Right because there's the same number here, same number here, same number here. But because this is a smaller area, they were close together. What about up here?
Student: Sparse.
Teacher: Yes, spread out – sparse. We've found a new word. What's the new word for today? So they're more spread out.

When comparing densities of the data there is a dual comparison: for each quarter within one group and between the two groups. This dual comparison was not communicated in this episode, rather she focused only on the female box plot. In the second teaching episode the spreads at a global level were compared and therefore the dual comparison tended to be communicated to students in a vague manner. In her reasoning the distinction between comparing variability within and between box plots was not well articulated. Although she did not consider the shape of the distributions, such as symmetry or skewness, in these teaching episodes she did in another teaching episode involving a matching exercise between histograms and box plots. At no time did she consider whether the shape was expected or unusual.

Element 6: Sampling The sampling element is underpinned by the belief that the data have been sampled from a population and that an inference will be made about the underlying population distributions from the sample distributions. David Pratt (personal communication, 7 July 2005) observed that confusion existed between the teacher and students about the game being played in this element. The students believe the teacher is making inferences about the samples, which Pratt refers to as *game one*, whereas the teacher is attempting to make inferences about the populations from the samples, which Pratt calls *game two*.

In the sampling element consideration is given to the sample size of each group and its effect on any inferences, whether a repetition of the experiment would give rise to the same difference, and determining the population for which the inference is applicable. In the 2003 analysis of the student assessment data about half the students, on the basis of fairness, mentioned that the sample size of the two datasets being compared should be the same (Pfannkuch, 2005). In cognizance of this finding the teacher brought students' attention to the datasets from teaching episode one:

Teacher: Yesterday's graph …did you remember that one set had 20 in it and the other set had 30 in it – were we still able to make comparisons between those sets?
Student: But that's not very fair.
Teacher: Not very fair?
Student: Like, with males and females.
Teacher: Yes, it's not but it doesn't necessarily affect your conclusion.
Student: Oh cause with box and whisker it doesn't matter cause it's just percentages aye?
Teacher: That's right, good call. Okay, so with the box and whisker it doesn't matter so much, although if one set could be smaller like say 5 people and you had say

> another set which had 30 and you were comparing them – then you probably would make some mention on that okay. But if they're roughly the same that's fine, doesn't have to be exactly the same.

The student seems to understand the sample size effect from a proportional basis, but the teacher's point appears to be that small samples (n=5) are more variable than larger samples (n=30). If she was playing game two she would point out that it would be unwise to draw conclusions from the comparison of such samples, but instead she states that it would be noteworthy. There is considerable conflict in this interchange since the teacher does not seem to have resolved which game she is playing.

When discussing the plots in teaching episode two the teacher attempted to hypothesize what would happen if she did the study again. Two ideas appear to be present in her discussion. The first idea is "if I took another sample from the population would I get the same results?" and the second idea may not have been the teacher's intention but it is worth considering, "if I repeated this experiment again would I get the same results?"

> Teacher: Now I'm going to throw in one more idea that I hope will convince you. This is 20 people, 20 people here. Okay, do you reckon, if we went back to the same place and did the same study and got another different 20 people, 20 boys and 20 girls, do you reckon we'll get exactly that same graph?
> Student: No.
> Teacher: Do you think, but don't you think that the median will be just a little bit higher for the girls?
> Student: Yes.
> Teacher: Do you think? Is it possible that maybe the results will be the other way around if it was another 20 people?
> Student: Yes.
> Teacher: Okay, can you see that really, they're so close, that if you were to get another 20 people, that it might just come out the other way. And then maybe in Mrs. L's classroom, maybe they've got that dataset and the girls graph might be a little bit higher than the boys and they'll be saying oh yes, and here we are looking at these, we read them the other way. So we have a bit of a problem here, I've only done this study once, we only did it with 20 girls and 20 boys, and probably *if we repeated the experiment*, we would find that we would have slightly different results.

The two ideas, taking another sample and repeating the experiment, are distinct. The first idea centers on the resultant outcome if a different sample was taken from the population and hence game two is being played. The second idea is to consider the consequences if the experiment was repeated on the same people. The implication of this idea is that another source of variation, namely measurement errors, should be considered but the game being played with this idea is game one. Again the sampling element becomes muddied.

The conflict between making inferences about samples, game one, and about populations, game two, is further illustrated by a student who wanted a definite conclusion and the teacher's subsequent conclusion.

> Student: But couldn't you say, from the graph, that males do have a little bit higher IQ than females?
> Teacher: … We're writing our conclusion: *"Based on these data values we are not certain that males have higher IQ."* It's not certain okay. *"There is some evidence to suggest that males have higher IQ for these students."*

The teacher's first statement draws a conclusion about populations whereas her second statement draws a conclusion about the samples. When writing her second statement ("There is some evidence to suggest that males have higher IQ for these students."), she draws students' attention to who was studied:

> Teacher: I'm even going to write, Uni students, all right, because these are all University students. I mean if we were trying to find out some information about all men and for all women and their IQ then this study wouldn't be enough. We would want to do surveys of people who are older, who are younger, who have different types of jobs, males, females, who are from New Zealand, Australia, India, China, Japan, Scandinavia, right we want to do that with people from all over the place. So we want to be really careful.

Unfortunately this reference to "Uni. students" was a game one statement, which further confuses the situation. Enculturating students into looking at who was studied in the sample and then being careful about determining the population on which the results can be generalized is part of learning about inference space judgment. Such a judgment is only possible when sufficient information is known about the data, which was not the case in teaching episode one.

Furthermore, during her discussion in teaching episode one she mentioned that if the distributions overlapped she would be careful about making a claim that men earned more than women, whereas if there was no overlap she would make the claim. The lack of overlap in sample distributions could be an artifact of sampling variation and hence the indeterminacy of her sampling reasoning is continued.

Element 7: Explanatory In the explanatory element, the background to and findings from the investigation are considered by referring to one's own real-world knowledge. This contextual knowledge is used to check whether the findings make sense or whether other variables should be considered before venturing a conclusion or hypothesis about the situation under consideration. Before students can compare box plots they need to understand the origin of the dataset, where and how the data were collected, and how the measures were defined. In teaching episode two the teacher first of all engaged the students' interest by telling them about an interesting talk on brain development that she had recently heard. Secondly, she discussed the measures used and on whom the study had been conducted. Part of her conversation was:

> Teacher: What, where does this data come from?
> Student: United States.
> Teacher: It comes from the United States, okay. What else do we know about this data?
> Student: They're all right handed.
> Teacher: They're all right handed, good, so all the people in this survey were right handed. What else do we know about these people?
> Student: They're university students.
> Teacher: They're university students, okay. What age group are university students usually?
> Student: Twenties?

Such information sets the data in context and lays the foundations for drawing reasonable inferences from data. In comparison, in teaching episode one, the data came from a textbook and all that was known about the data was that they were collected from a local firm. Such a paucity of background information led the teacher to consider female and

male salaries in general, in an attempt to discuss whether the findings made sense with what they knew about the world.

> Teacher: Did anyone see the recent results on the average salaries for men and women? I remember seeing something on the news about that. I think it was to do with people who work for the government and public service and that includes people like teachers, nurses, policemen, officials who work in government departments – everyone who gets paid by the taxpayer if you like, right, they did a survey to have a look at who earned more – men or women and they found that it appeared that men earned a little bit more than the females. …

Later, the discussion considered whether there was an alternative explanation for the findings rather than gender being the discriminating factor for salaries:

> Teacher: Like, teaching for example, whether we're female or male maybe doesn't affect how much we earn, but maybe it affects things like –
> Student: What position you're in.

Such a discussion enables more variables to be considered before making an inference based on given data. Thinking of confounding variables and alternative explanations for findings are part of the argumentation with data, and more information on this dataset, together with other relevant data, could have provided a richer exploration.

Element 8: Individual Case When reasoning from distributions, observations which appear to be outliers are inspected as individual cases to determine whether they are part of the dataset or are errors and can be corrected or removed. Because box plots are not drawn with outliers at this Year level and dot plots were not kept under the box plots this element was not articulated. However, the teacher did reason with individual cases when she was arguing from a hypothesis generation element:

> Teacher: This person here, this woman here, earns $605, she earns way more than this male here $257.

The comparison of individual's earnings between the datasets is a method of argumentation based on particular instances that is used to illustrate that definitive statements cannot be made for all cases.

3.2. THE MODERATING ELEMENTS OF REASONING

The moderating elements of reasoning, evaluative and referent, serve two distinct supporting functions in the reasoning process. The evaluative element's function is to support the reasoning process by qualitatively judging the strength of the evidence provided by an element and then weighing up that evidence towards making a decision about whether there is a real difference between the two groups under consideration. The referent element's function is to ground and maintain the reasoning process within contextually-based data, since the box plot is a representation that compresses and obscures information.

Element 9: Evaluative As each of the eight elements is considered, the evidence provided by that element is described, assessed on its strength. and weighed up in the process of making a judgment on the data. For example, for teaching episode one, a description would be "the male graph is *higher* than the female graph," whereas the

strength of the evidence is conveyed by "the male graph is a *lot higher* than the female graph." Weighing the evidence is conveyed by statements such as "*even though the graphs overlap* these data suggest males on average earn more than females." In teaching episode two, the teacher describes, assesses and then weighs all the evidence she has accumulated. The language of this evidence is italicized.

> Teacher: Now I know the numbers are different, the males are *bigger* than the females, but *it's not that different*, it's not like one's 100 and the other's – you know? So, it's another contributing thing, men's stuff is *more spread out*. But *it's not massively different*, especially when you see it on the graph, you know, *it's not that different*, can you accept that? Okay, so at the moment, I've *got some conflicting kind of information*, right median – females are *more* clever, but when I look at the whole graph, the whole graph's a *bit more higher* for males. They're a little bit different in their spreads but you know, so *I'm still not ready to say yes* males have got a higher IQ than females.

The weighing of evidence involves qualitative judgments and a subtle use of language to convey how a decision is being reached. Since informal inferences are being made it may be hard for students to determine in inconclusive situations what evidence is taken notice of by the teacher when making a decision (see Figure 2(c)).

Element 10: Referent When the teacher is comparing two distributions represented as box plots in a symbolic system, then reasoning from this symbolic system necessitates a constant reference to other systems. The box plots are constantly being decoded in a back-and-forth switching between the visual symbol system and the concepts and ideas to which it refers. For example, the teacher in the spread element of reasoning decodes the visual system, a rectangular box divided by a line with a whisker at each end, when she imagines a quarter of the females standing in each section. Such an imagining, with some females standing closer together than others, is a switch to another reference system or another representation of the box plot. Her main referents were the context or the statistical measures the symbol system was portraying. For example, she said "*female* graph is higher," "*male earnings* are higher, or "*female median* is higher." Sometimes her referent was the imagined underlying distribution of the data, "this central box here gives me the *middle 50% of the data*." Her language did not refer very often to the underlying plots, which had been summarized by the box plots. Furthermore, her referents to the data plot for her justifications in the written conclusions (Figures. 1(d), 2(c)) seem to be insufficient.

3.3. LIMITATIONS

There are two main limitations to this research. First, the study has only captured one teacher learning to communicate her reasoning from box plot distributions. Second, one researcher categorized the elements and hence there is no triangulation from independent sources, although the teacher did take the opportunity to assess the interpretation. Because there seemed to be no account of how to draw informal inferences from the comparison of box plots at an introductory level, an action-research method, where learning to reason occurred in and from practice, was deemed appropriate. Hence, the research can only offer some insight into possible ways teachers and students could be expected to reason informally and into possible pitfalls in the reasoning process. The research also makes the case for developing sampling reasoning concepts and keeping data with the box plots, but again this is based on one teacher's reasoning. Therefore the

discussion that follows draws on the literature from students' reasoning to support some findings but remains speculative in terms of teachers' reasoning.

4. DISCUSSION

Informal inference should be stimulating intuitive foundations for formal inference. Making informal inferences based on distributions alone is not the usual statistical practice and hence the teacher in this study should be viewed as a learner in a new situation struggling to convey the messages in data. Indeed Biehler (1997, p. 176) stated that "there are profound problems to overcome in interpreting and verbally describing statistical graphs that are related to the limited expressability of complex quantitative relations by means of a common language" and that researchers need to become more aware of the difficulties.

The key finding from this research is the proposal of a descriptive model (Figure 3) of reasoning from box plots. The model is complex and is the beginning of an exploration into the elements of reasoning that could be considered when structuring teaching towards formal inference. The elements, hypothesis generation, summary, shift, signal, spread, and individual case, have been described by other researchers. The shift element could be incorporated into another element and may not be as important as the others in the reasoning process. The summary element, especially the comparison of equivalent five-number summaries, may be considered unimportant but nevertheless such reasoning does exist and the purpose is to document all the types of reasoning invoked. The main findings from this research, however, are the description of ways in which the elements of sampling, explanatory, evaluative and referent are also part of a reasoning process that leads towards formal inference. Each element will now be discussed.

The hypothesis generation element is an aggregate-based reasoning approach that Konold, Pollatsek, Well, and Gagnon (1997) and Ben-Zvi (2004) believe is essential if students are to reason about trends and patterns in distributions. The teacher's discourse incorporates such notions. Her reasoning also highlights the link between the nature of the representation and the nature of the reasoning, particularly in the summary and shift elements. The teacher's reasoning is led intuitively towards comparing the five-number summary boundaries, a facet of reasoning Biehler (2004) and Friel (1998) noticed in their students. The shift element documented by Biehler (2004) appeared to be intuitively inherent in the visual nature of the representation and hence in the teacher's reasoning.

Reasoning with measures of center is expressed by the teacher in the signal element. Both Bakker (2004) and Biehler (2004) report that the median as a representative value of a distribution is difficult to develop. When Bakker (2004) and Konold et al. (2002) searched for an alternative notion of center, they hypothesized that students' intuition of middle group or modal clump could support the development of center as being a characteristic of the distribution. The teacher does use the middle 50% of data as an intuitive device for the signal. Also the nature of the representation leads to this type of argumentation.

Within the spread element of reasoning by the teacher, two comparisons are evident: comparing the densities within one box plot and comparing the densities between the two box plots. Such a discussion was not clear to students nor was the purpose of the discussion of how comparing spreads helped in making an inference. Biehler (2004) noted that his students did not comment on spread differences. Another problem with the spread element, which is closely aligned to the nature of the representation, is how concepts can be built up for viewing spread as a dispersion from the median, which according to Bakker (2004) is a big transition. When the teacher compared the overlap of

the boxes with drawn lines, she was taking into account some of the variability. The question is whether such a comparison could be conceived as an intuitive beginning for confidence interval ideas for true population medians and for viewing spread as dispersion from the median.

The sampling reasoning element is presented by the teacher via thought-simulations rather than by empirical simulations in which students could actually experience the variability of samples drawn from populations (Pfannkuch, 2005). Both verbalization and experience of sampling behavior are necessary if teachers and students are truly to grasp the nature of sampling reasoning. Moreover, this element is key to bridging students towards formal inference. The game to be played is game two whereby the reasoning involves making inferences about populations from samples, not making inferences about samples, game one. The teacher did not resolve which game she was playing and therefore a large part of inferential reasoning eluded her and her students. In order to play game two, activities, such as "growing a sample" (Bakker & Gravemeijer, 2004), bootstrapping (Finzer, personal communication, 7 July 2005), and experiencing and building concepts about sampling behaviour (Pfannkuch, 2005) could assist in developing her and her students' sampling reasoning. Reasoning about samples also includes how the sample was selected and sample size (Watson, 2004). Although the teacher referred to sample size, she did not discuss how the sample was selected as that information was not presented as background information, but she was careful in acknowledging who was sampled and on whom she could draw an inference. In other words she paid attention to inference space judgment.

For the explanatory element a way of perceiving her reasoning is to consider that the distributions are a statistical model of a real world situation. Since contextual knowledge is essential for seeing and interpreting any messages in data, a continuous dialogue should exist between the statistical models and the real world situation. Hence features seen in data produce queries about context, which in turn suggest questions for the data (Wild & Pfannkuch, 1999). This continuous shuttling between the contextual and the statistical is present in the teacher's reasoning. Her choice of learning task for teaching episode one, however, illustrated how lack of background information about data leads to speculation about the data rather than further exploration. Context is used by the teacher as an integral part of the interrogation of data, as a factor in determining whether confounding variables are present, and for determining whether there are alternative explanations for the findings. Friel et al. (2001, p. 140) also highlight that the contextual frame of data is necessary for comprehending and making judgments on graphs although it increases "the number of elements to which the graph reader must attend."

The moderating elements of reasoning, the evaluative and referent elements, act as anchors for weighing the evidence and for interpreting an abstract box plot representation respectively. The evaluative element includes making a judgment by comparing distributions and is alluded to by Friel et al. (2001) in their suggested taxonomy of judgment tasks. For actually making an informal inference this element is critical. Qualitative judgments on the whole must be made to ascertain whether one is prepared or not prepared to state Group A is greater than Group B, on average. Within each of the eight elements of reasoning, the teacher is continually making qualitative and sometimes quantitative statements as a prelude to weighing the evidence. Weighing the evidence is a matter of opinion, can be subjective, and rests on experience with data. The students' lack of experience and seemingly innate need for a definite conclusion (see sampling element dialogue) may militate against realizing that in statistics findings may be inconclusive.

Bakker and Gravemeijer (2004) consider referents as being essential for instructional design. The symbol system, the box plot, is a new representation, and students may need

to interpret it with a better-known system such as a dot plot where individual data are identifiable (Carr & Begg, 1994). Friel et al. (2001, p. 139) also note that a "major component of the graph reader's interpretation process is relating graph features to their referents." The teacher's referents are many-fold, each acting to place the abstract representation into a context as well as to imagine the data underneath the box plots. For someone with her experience there may be no problem in imagining the plot underneath, but for the students the abrupt transition from the stem-and-leaf plot to the box plot may have been too fast (Pfannkuch, 2006).

According to Moore (1990) and Wild and Pfannkuch (1999), variation is at the heart of statistical thinking. All the elements are underpinned by variation as it is noticed, dealt with, measured informally, and explained. Or as Finzer (personal communication, 7 July 2005) more succinctly stated, "distribution reasoning is the recognition and utilization of patterns in variability." Reasoning about distributions is more than reasoning about shapes (Bakker & Gravemeijer, 2004), it is about decoding the shapes (Friel et al., 2001) by using deliberate strategies such as the proposed model (Figure 3) to comprehend distributions. Furthermore, there is a weighing of evidence to form an opinion on and inference from the information contained in the comparison of distributions. Such informal decision-making under uncertainty requires qualitative judgments, which would seem to be much harder than the quantitative judgments of statistical tests.

The analysis of one teacher's reasoning from box plot distributions contributes to the research base by enhancing understanding of the reasoning processes, and raising issues about the links to formal inference, the nature of the game being played, and instructional practice. The model (Figure 3) demonstrates the richness of verbalization necessary for communicating ideas and concepts from box plot distributions, and builds on other research findings. Thus the model begins to propose a coherent framework for the nature and type of informal inferential reasoning that might be addressed when teaching students how to reason when comparing box plot distributions.

5. IMPLICATIONS FOR RESEARCH AND EDUCATIONAL PRACTICE

More research work is needed on designing instruction and building teachers' and students' concepts and reasoning about distributions towards formal inference. Research is also needed on developing teachers' and students' sampling conceptions in terms of learning to reason about populations from samples using informal inference. Since this research is based on one teacher's reasoning in a non-technological environment, there may be other reasoning elements necessary for informal inference. The challenge for future research is to move towards a prescriptive model of reasoning from box plot distributions. Such a model could specify how the different reasoning elements could be woven and sequenced together during instruction and exemplify how the elements contributed towards the development of formal statistical inferential reasoning.

At the teaching level, the implications from this research suggest that developing teacher and student talk on how to communicate ideas and on concepts represented in distributions are essential. The model suggested by this research has now been used as a guide in developing teacher reasoning and for writing down how to reason from box plots. Instruction, however, needs to adopt a gradual transition approach from dot plots to abstract box plots to improve the referent element of reasoning and to build the sampling reasoning element through giving teachers and students opportunities to experience sampling behavior. Such an opportunity was taken by the teachers in this project in 2006.

Hill, Rowan, and Ball (2005) believe that teachers' mathematical content and pedagogical content knowledge are linked to student achievement and that improving

teachers' mathematical knowledge will improve students' understanding. Teachers and researchers need to collaborate to develop a coherent, deeper conceptual approach to the learning of statistics. A research agenda should be implemented since the current situation in teaching and assessment requires teachers and students to make informal inferences from the comparison of distributions. Without an underlying research base on informal inference and reasoning from distributions, this situation may lead to some unforeseen consequences in later years of schooling.

REFERENCES

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools.* Utrecht, The Netherlands: CD-ß Press, Center for Sci. and Math. Education.

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education (IASE) Roundtable*, Lund, Sweden, 28 June-3 July 2004, (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.
[Online: www.stat.auckland.ac.nz/~iase/publications.php]

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ball, D., & Cohen, D. (1999). Developing practice, developing practitioners. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco: Jossey-Bass publishers.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42-63.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_BenZvi.pdf]

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 169-190). Voorburg, The Netherlands: International Stat. Inst.
[Online: www.stat.auckland.ac.nz/~iase/publications/8/14.Biehler.pdf]

Biehler, R. (2004, July). *Variation, Co-Variation, and Statistical Group Comparison: Some Results from Epistemological and Empirical Research on Technology Supported Statistics Education.* Paper presented at the 10th International Congress on Mathematics Education, Copenhagen.

Carr, J., & Begg, A. (1994). Introducing box and whisker plots. In J. Garfield (Ed.), *Research papers from the Fourth International Conference on Teaching Statistics (ICOTS 4)* Marrakech, Morocco, July, 1994. Minneapolis, MN: University of Minnesota.

Friel, S. (1998). Comparing data sets: How do students interpret information displayed using box plots? In S. Berenson, K. Dawkins, M. Blanton, W. Coulombe, J. Kolb, K. Norwood, & L. Stiff (Eds.), *Proceedings of the Twentieth Annual Meeting, North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 365-370). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Friel, S., Curcio, F., & Bright, G. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124-159.

Hill, H., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371-406.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 151-168). Voorburg, The Netherlands: International Statistical Institute.
[Online: www.stat.auckland.ac.nz/~iase/publications/8/13.Konold.pdf]

Konold, C., Robinson, A., Khalil, K., Pollatsek, A. Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society*, Cape Town, South Africa, July, 2002 [CD-ROM.] Voorburg, The Netherlands: International Statistical Institute.
[Online: www.stat.auckland.ac.nz/~iase/publications/1/8b2_kono.pdf]

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

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267-294). Dordrecht, The Netherlands, Kluwer Academic Publishers.

Pfannkuch, M. (2006). Informal inferential reasoning. *Proceedings of the Seventh International Conference on Teaching Statistics: Working cooperatively in statistics education*, Salvador, Brazil, July, 2006 [CD-ROM.] Voorburg, The Netherlands: International Statistical Institute.
[Online: www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Pfannkuch, M., Budgett, S., Parsonage, R., & Horring, J. (2004). *Comparison of data plots: building a pedagogical framework*. Paper presented at ICME-10, Denmark, TSG11: Research and development in the teaching and learning of probability and statistics.
[Online: www.stat.auckland.ac.nz/~iase/publications/11/Pfannkuch.doc]

Pfannkuch, M., & Horring, J. (2005). Developing statistical thinking in a secondary school: A collaborative curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education (IASE) Roundtable*, Lund, Sweden, 28 June-3 July 2004, (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.
[Online: www.stat.auckland.ac.nz/~iase/publications/rt04/5.1_Pfannkuch&Horring.pdf]

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Watson, J. (2004). Developing reasoning about samples. In D. Ben-Zvi and J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, *67*(3), 223-265.

MAXINE PFANNKUCH
Department of Statistics
The University of Auckland
Private Bag 92019
Auckland, New Zealand