THE LANGUAGE OF SHAPE

Pip Arnold[1,2] and Maxine Pfannkuch[2]

[1]Cognition Education Ltd, [2]The University of Auckland

parnold@cognition.co.nz, m.pfannkuch@auckland.ac.nz

Students often struggle with describing the shape of different data distributions as they are distracted by the noise and do not "see" the signal. Their attention is drawn to the actual outline of the distribution rather than an inferred distributional shape. In this paper we describe part of an instructional sequence for learning about shape starting with large data distributions. The instruction was trialled in a year 10 class (age 14) and included a focus on developing the language of shape for describing distributions and identifying key features for description. Responses from pre- and post-tests are briefly discussed and a proposed framework for describing distributions is presented.

Keywords: Secondary students; Statistics education; Describing distributional shape

INTRODUCTION

> The power of statistical data analysis lies in describing and predicting aggregate features of data sets that cannot be noted from individual cases (Bakker, 2004, p. 100).

In a New Zealand national assessment students in year 11 (age 15) are expected to be able to undertake a statistical investigation about a comparison situation. The assessment requires students to (in brief): pose an appropriate comparison investigative question; select and use appropriate display(s); give summary statistics; discuss features of distributions comparatively, such as shape; and communicate findings in a conclusion. For many years teachers have struggled with exactly what describing the shape of a distribution means and recent research on informal inferential reasoning identified describing shapes of data distributions as an area where students demonstrated impoverished reasoning (Pfannkuch, Arnold, & Wild, 2011). Discussions were held with overseas experts and a clear solution was not evident, though fledgling ideas existed. These ideas were developed into activities to explore several aspects of distribution including the language of shape, making predictions and building a contextual knowledge base about shape.

LITERATURE REVIEW

Over the last ten years there have been a number of research projects with a focus on distribution and students' reasoning about distribution; for example, the Freudenthal Institute team (Bakker, 2004; Bakker & Gravemeijer, 2004), the Nashville team (McClain, 2005; McClain & Cobb, 2001), and the 2005 Fourth Statistical Reasoning, Thinking and Literacy Research Forum focused on reasoning about distribution. Five themes emerged from the research: (1) the notion of distribution; (2) measures of centre; (3) shape of distributions; (4)

predicting distributions; and (5) contextual knowledge. This paper will focus on the shape of distributions and describing distributions.

Distribution is a multi-faceted notion involving centre, spread, skewness, shape and density (Bakker, 2004; Ben-Zvi & Amir, 2005; Konold, Higgins, Russell, & Khalil, 2004; McClain, 2005; Pfannkuch, 2005; Reading & Reid, 2006). Students need to consider measures of centre, measures of spread, where the majority of data values are in relation to extreme values, and how density and skewness provide detail about shape when viewing distributions. It is this global reasoning, the coordination of these ideas that makes distribution a complex notion that students find difficult (Ben-Zvi & Arcavi, 2001; delMas, Garfield, & Ooms, 2005; Hancock, Kaput, & Goldsmith, 1992; McClain & Cobb, 2001).

Describing shapes of distributions has had fleeting mention, with Bakker (2004) providing the only real depth in work on shape. Despite the relative superficial exploration of shape there are some starting points to consider. Firstly, the type of graph used to display the data has a major influence on students' ability to perceive shape. For example, box plots and even histograms at earlier ages can prove a problem for students to use as they are too abstract and the actual data cannot be seen (Bakker, 2004; Friel, Curcio, & Bright, 2001). Dot plots on the other hand provide an initial starting point for students to explore shape along with simple case-value bar graphs (Bakker, 2004; delMas et al., 2005). Pfannkuch (2005) suggests that dot plots and stem-and-leaf plots can provide a strong basis for interpreting and understanding distributions and students can transition from them to box plots. Most critically displays used should allow sense to be made of the information with as much ease as is possible (Friel et al., 2001). Secondly, Bakker (2004) suggests that single univariate distributions are a good starting point, but cautions that students can initially assume that all distributions are symmetric if only this type are selected. Students' thinking can be challenged by deliberately choosing distributions that are skewed as well as symmetric (Bakker, 2004; delMas et al., 2005; Makar & Confrey, 2005; Rubin, Hammerman, Puttick, & Campbell, 2005). Linked to this is providing many opportunities for students to recognise and understand the direction of a skew (delMas et al., 2005), which is also a problem for college level students. Descriptors of shape include uniform, normal, skewed to the right or left (Bakker, 2004) and normal, skewed, bimodal or uniform (delMas et al., 2005), with early student ideas describing the data in terms of low, average and high values and naming shapes using pyramid, semi-circle and bell shaped (Bakker, 2004). Thirdly, shape helps to develop meaning for mean, spread, density and skewness (Bakker, 2004; Rubin et al., 2005) and connections between measures of centre and shape can be made (Konold & Higgins, 2003; Rubin et al., 2005). Finally, Bakker (2004) found that too small a sample size, unsuitable scaling and lack of context were problems when trying to identify the shape of distributions.

An end goal is that students are able to describe sample distributions as part of the statistical enquiry cycle to answer an investigative question about a population. The research questions for this paper are: What shapes do year 10 students (age 14) realise from data distributions? and What descriptions of distributions are year 10 students capable of producing?

THEORETICAL FRAMEWORKS

Two theoretical frameworks were considered in the analysis of student responses in pre- and post-tests. Bakker and Gravemeijer (2004) proposed a structure (Fig. 1) for analysing the relationship between data and distribution. They said that students as novices typically see individual values and use these to find values such as the median, range or quartiles, but that this does not mean they are seeing the median, for example, as representative of a group.

distribution (conceptual entity)			
centre mean, median, midrange, …	**spread** range, interquartile range, standard deviation, …	**density** (relative) frequency, majority, quartiles	**skewness** position majority of data
data (individual values)			

Figure 1. Between data and distribution (Bakker & Gravemeijer, 2004, p. 148)

Ben-Zvi, Gil and Apel's (2007) informal inferential reasoning (IIR) theoretical framework provides cognitive aspects that relate to distribution – reasoning about variability (spread, density), distributional reasoning (aggregate views, pattern and trend, hypothesis and prediction, individual cases, outliers), reasoning about signal and noise (centre, measures, modal clumps, summary), contextual reasoning (interpretation, alternative explanations) and graph comprehension (decoding visual shapes).

METHODOLOGY

The research method follows design research principles (Roth, 2005) for a teaching experiment in a classroom. In the preparation and design stage the first author developed the teaching and learning materials to use in the teaching experiment in conjunction with the classroom teacher, considering relevant literature. In addition, purposefully built into the teaching and learning sequence were activities with a focus on shape prediction and building a "library" of knowledge around contexts and shape but these are not reported in this paper. Both the classroom teacher and first author were involved in the implementation of the activities in the teacher's year 10 class. Following each lesson there was reflective discussion and adjustments were made as needed to the learning trajectory.

The 29 students in the class were above average in ability and from a mid-size (1300), multicultural, mid socio-economic inner city girls' secondary school. Students were given a pre and post-test, the lessons were videotaped and student work was photocopied. A group of six girls were observed specifically as well as the teacher led whole class discussions. The six girls also had pre and post-interviews about their responses to their tests.

The retrospective analysis for this paper focuses on the development of students' use of the language of shape and their descriptions of distributions. The learning activities were designed to support students' understanding of these two aspects. The activities built on work previously undertaken in an informal inferential reasoning project (Pfannkuch et al., 2011). They also included new thinking as we considered the bigger picture of what we were trying to achieve. The new/updated activities were based on the themes that emerged from the literature: in particular they focused on the language of shape, making predictions and building a contextual knowledge base for the sorts of variables that have symmetric, skewed or uniform distributions. Unpacking students' existing contextual knowledge and misunderstandings were key ingredients in predicting distributions.

This paper focuses on lessons 2-4 of a 16-lesson unit on statistics. Lesson one was a review. Lesson two had a focus on seeing and describing shape and involved developing the language for shape of distribution descriptors, sketching shapes from graphs, grouping similar shaped graphs, and matching shape descriptors to groups of graphs. Shape of distributions was a big idea in the lesson. Lesson three had a focus on linking shape and context and involved making predictions of the graph from contexts, matching contexts to graphs, and starting to develop a "library" of similar shaped graphs. Big ideas in this lesson were shape of distributions, predicting distributions and contextual knowledge. Lesson four focused on using the language of shape to describe distributions and involved sorting graphs according to shape of distribution and starting to describe distributions. The big ideas in this lesson were the notion of distribution, shape of distributions and contextual knowledge. The first author (FA) taught lessons three and four as the teacher was ill.

TEACHING ACTIVITIES

The three lessons described demonstrate how student-generated concepts, ideas, and language were gradually transformed towards a statistical approach.

Lesson 2: Seeing and describing shape

The students firstly sketched the shape of 15 data distributions that were briefly shown using a PowerPoint presentation. Secondly, the students grouped the sketches of the graphs into similar shapes and used their own language to describe the shapes in each group. At this point the teacher asked about the number of groups they had made. For example:

Teacher: four groups, what were they based on?

Student: sloped to the left, and sloped to the right, symmetric ones

Teacher: so you have sloped to the left, sloped to the right, symmetric and what was your other group?

Student: you know [gestures with hand – up, across and down] it is even on the top

Teacher: even on the top, so let's see, symmetrical, some sloped to the left, slope to the right, other one was…[Various student responses with "flat top" being the loudest.]

The teacher used these four group headings – symmetrical, sloped to the left, sloped to the right and flat top – as a starting point. The class then sorted the graphs into one of the four groups (Fig. Figure 2). Finally the students were introduced to the statistical language used to describe shapes and were asked to match these words to their graphs. Intuitively the students re-grouped the graphs according to symmetry, symmetric or not symmetric, splitting the skewed into two groups (left and right) and the symmetric into two groups (uniform and other). Interestingly modality was not used for grouping.

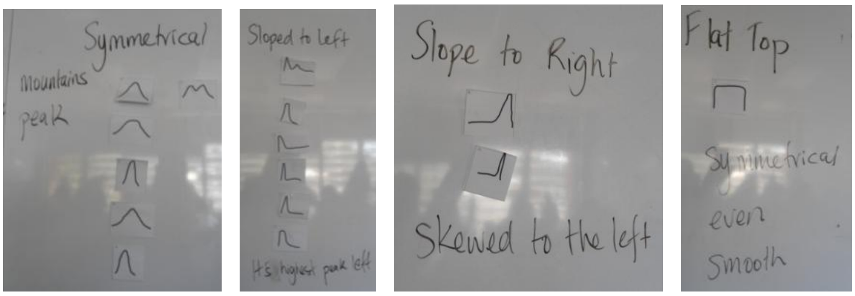


Figure 2. Four shape groups with graphs and additional ideas

Lesson 3: Linking shape and context

In order to get students to think about how context and shape were linked, they were given 15 contexts without the graphs and asked to sketch a shape for these contexts with some possible values. Discussion justifying shapes for particular contexts followed. Students were then given the actual dot plots of the contexts and they matched these plots to the context. The graphs were sorted again into the four groups (symmetric, sloped to the left, sloped to the right, flat top) and each group was labelled using appropriate statistical terminology – symmetrical, right skew, left skew and uniform, including a discussion around why and which way the skew was recognised. At this point the distinction between unimodal and bimodal was made.

FA teacher: These are the graphs yesterday that you said were symmetric, and I've moved this one out to the bottom. Why do you think I have done this?

Student: Because it's bimodal.

FA teacher: Because it's bimodal. So these are symmetric, and unimodal, which means that they have one bump or one peak. So they have one mode, or peak and this one here is symmetric and bimodal because it has two peaks.

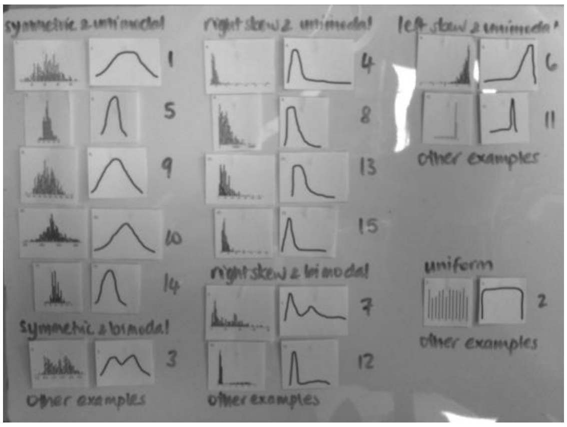


Figure 3. Final collation of shapes into four groups with modality distinction

From this brief conversation the way to sort the shapes became clear – sort by symmetry and then by modality (Fig. 3). The shape descriptors developed from the way the students intuitively sorted the graphs. The students did not make a separate group for bimodal, as the research team did. They sorted into four groups and then split three of the groups by modality.

Lesson 4: Using the language of shape to describe distributions

The students firstly classified some more data graphs by shape and added these to their growing "library" of shapes and contexts (note in Fig. 3 the label other examples). The next activity involved starting to describe distributions. The FA teacher facilitated a class discussion on key features for describing graphs. They were given the challenge that if they had to draw the graph from the description, what information would they need. Shape was a given, but a number of other features surfaced. A few excerpts from the discussion are:

FA teacher: So if I was going to describe this graph what other things might I want to describe about it?

Student: The range.

FA teacher: What other things would be important?

Student: Its highest point.

FA teacher: What are we calling that highest point?

Student: The peak. …

FA teacher: What else might we want to talk about? What makes that graph (points to number 3) different to say number 14 (see Fig. 3)?

Student: The amount of peaks [modality].

In the discussion the features that the students suggested included: target population, variable, units, general shape sketched, overall shape, modality, peaks, range, median and mode. A further conversation in the same lesson where the focus was on describing one of the right skew graphs additional features surfaced: clustering density, majority, modal group and describing shape in terms of parts of the whole.

PRE AND POST-TEST WRITTEN RESPONSES

Student pre- and post-test responses were analysed to see if their ability to describe distributions had improved over the course of the statistics unit. In one of the questions students were asked for each of three situations (see Fig. 4) to sketch the shape of the distribution of the variable and to write two statements about the distribution of the variable.

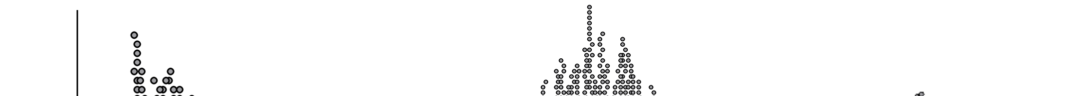


Figure 4. (a) All Blacks' (NZ rugby team) scores in test matches 2005-2010; (b) heights of NZ Year 5-10 students; (c) heights of Tokoeka Kiwis (NZ native bird)

The SOLO taxonomy (Uniservices asTTle team, 2008) was used as a basis for grading student responses. The particular descriptors aligned to each question were developed through a process of moving between the literature, in-class observations and student responses. In brief the descriptors for grading the student responses are: **no response** (NR-0);

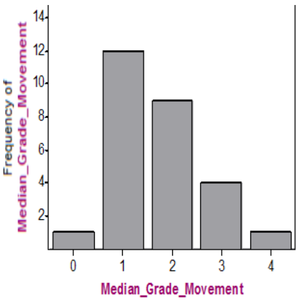
pre-structural (PS-1) – context and/or evidence missing; **uni-structural** (US-2) – gives one correct piece of evidence in simple context OR multi-structural evidence without any context; **multi-structural** (MS-3) – identifies a simple context and correctly describes two features OR relational evidence without any context; **relational** (R-4) – identifies the context, connects the context, and correctly describes the overall shape and at least two other features; **extended abstract** (EA-5) – identifies the context, connects the context throughout the description, correctly describes the overall shape and at least three other features and may include some explanation or interpretation of results to the context (see Fig. 5(c) for an example of an extended abstract response).

Post-test responses

Pre-test responses

	NR-0	PS-1	US-2	MS-3	R-4	EA-5	Total Pre
NR-0	0	0	0	1	0	0	1
PS-1	0	0	2	7	1	1	11
US-2	0	0	0	9	2	2	13
MS-3	0	0	0	1	1	0	2
R-4	0	0	0	0	0	0	0
EA-5	0	0	0	0	0	0	0
Total Post	0	0	2	18	4	3	27

(a) Pre- and post-test results



(b) Median grade movement from pre- to post-test

The distribution of the heights of Tokoeka kiwis is approx symmetrical and bimodal. The heights range from 35-43cm. The middle tokoeka kiwi height is 39cm. The heights peak at around 36.5 and 40cm. The heights are tightly grouped in two groups one between 36-39cm and another between 39-42cm. These two groups might mean the two different genders.

(c) Extended abstract response example for plot in Fig. 4(c)

Figure 5. Pre- and post-test results for one assessment question

The median grade across the three situations was used to represent the students' overall grade. These are summarised in Figure 5(a). In the pre-test the highest median grade was multi-structural with two students achieving this. In the post-test three students achieved at extended abstract and all but two students reached at least a multi-structural level. This means that the students could identify the context and describe at least two features. A lot of these students actually described more than two features, but they failed to make the broader link to the context, which was required to show relational thinking. The biggest movements were from students who scored 0-2 in the pre-test, perhaps indicating that acquisition of language and knowledge for describing distributions assists students. Figure 5(b) shows the median difference between students' pre- and post-test scores. The students made a significant improvement (*P*-value≈0) in their median scores from the pre- to post-test question and on average increased their median grade by 1.7 points (95% C.I.= [1.34, 2.07]).

CONCLUSIONS

> "Distribution" is another fundamental given of statistical reasoning. I can find a great deal written about specialised usages and definitions of "distribution" but almost nothing about "distribution" itself as an underlying conceptual structure (Wild, 2006, p. 10).

The research questions were: What shapes do year 10 students (age 14) realise from data distributions? and What descriptions of distributions are year 10 students capable of producing? The year 10 students in this study intuitively sorted the data distributions into four groups, symmetric, right and left skew and uniform. These groups were further refined using a modality distinction. The classification realised by the students was based on what they noticed as they sought to group the graphs by shape. The teacher acknowledged student language and introduced appropriate statistical language which was connected to their four groups. These year 10 students appear to have the capacity to write thorough descriptions of data distributions. Further work and teacher modelling is needed to move students to a relational level where they can see the significance of parts of the whole description and intertwine context throughout the description.

Distribution is a complex notion. During the retrospective analysis phase, when student pre- and post-test responses were analysed, the two frameworks (Bakker & Gravemeijer, 2004; Ben-Zvi et al., 2007) that had previously been considered were found to only provide part of the picture. These frameworks needed to be linked with a specific focus on the underlying conceptual structure of distribution. Combining the two frameworks led to our new proposed framework, the Distribution Description Framework (DDF), for thinking about, exploring and describing distribution. The DDF (Fig. 6) is organised by: (1) *overarching statistical concepts* that underpin distribution (2) *characteristics of distribution*, and (3) the *specific features* that are used when describing distributions.

Overarching statistical concepts	Characteristics of distribution	Specific features measures/depictions/descriptors
Contextual _knowledge_	_Population_	_Target population_ (e.g. NZ Yr 5-10 students) Other acceptable population (e.g. Yr 5-10 students)
	Variable	Variable Units
	Interpretation	_Statistical feature described in contextual setting_ (e.g. interpreting right skew as very few high test scores, with most test scores between 20-50 points)
	Explanation	_Possible reason for a feature_ (e.g. bimodal due to gender for kiwi data)
Distributional	Aggregate view	General shape sketched Hypothesis and prediction
	Skewness	Position of majority of the data
	Individual cases	_Highest and lowest values_
Graph Comprehension	Decoding visual shape	Overall shape *_Parts of the whole_ (splitting the distribution into parts and describing the parts as well as the whole) _Modality_
	Unusual features	Gaps Outliers
Variability	Spread	Range, inter-quartile range *_Interval for high and/or low values_ (may be describing a tail)
	Density	_Clustering density_ Majority (mostly, many) Relative frequency
Signal and noise	Centre	Median, mean
	Modal clumps	_Peak(s) (local mode)_ _Modal group(s)_

Figure 6. Distribution Description Framework for year 10 (*indicates part of feature listed)

Ben-Zvi, Gil and Apel's (2007) cognitive aspects from their IIR theoretical framework – reasoning about variability, distributional reasoning, reasoning about signal and noise, contextual reasoning and graph comprehension – were used to inform the overarching statistical concepts for distribution descriptions. Bakker and Gravemeijer's (2004) characteristics of distribution – centre, spread, density and skewness – formed the backbone, with Pfannkuch, Regan, Wild and Horton's (2010) ideal data-dialogue providing further characteristics and features to supplement those listed in the IIR theoretical framework. The result of the analysis of student pre- and post-test responses and in-class observations in this research provided the additional characteristics and features noted in Figure 6 in italics.

Collectively these sources of data and ideas build a richer picture of the possible features that may be present in a particular distribution. While some aspects will be true and relevant in all descriptions (e.g. variable, overall shape), others (e.g. clustering density, mode) will depend on the data and whether or not they are relevant in the description. When students are describing statistical distributions they need: (1) to invoke contextual knowledge, (2) to know what relevant characteristics of distributions they can actually see in the plots and therefore describe, and (3) to be explicit about the evidence for specific features. In other words, students need to be able to identify which features are evident in a particular plot, name and provide evidence (values) for the features and to interlace these with contextual information such as the population, variable and units.

Bakker and Gravemeijer's (2004) framework appears to be about data distributions. In this study the data distribution is conceived as a sample distribution and therefore more concepts come into play. At year 11 students are introduced to new concepts such as sampling variability, sketching inferred shapes and comparing sample distributions. This means that the DDF would be extended with students co-ordinating more ideas. The DDF would expand again in senior secondary where students start to consider distributions of statistics. Similarly, the DDF can be modified to support student progressions at lower curriculum levels. We believe our DDF has the potential to inform curriculum developers, researchers and teachers as they introduce students to the conceptual structure underlying distribution. Further research is needed both above and below the level reported here to ascertain what is appropriate for students at the different levels.

References

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht: Freudenthal Institute.

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.

Ben-Zvi, D., & Amir, Y. (2005, 2-7 July). *How do Primary School Students Begin to Reason about Distributions?* Paper presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4), Auckland, New Zealand.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35-65.

Ben-Zvi, D., Gil, E., & Apel, N. (2007, August 11-17). *What is hidden beyond the data? Helping young students to reason and argue about some wider universe.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

delMas, R., Garfield, J., & Ooms, A. (2005, 2-7 July). *Using Assessment Items to Study Students' Difficulty Reading and Interpreting Graphical Representations of*

Distributions. Paper presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL), Auckland, New Zealand.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124-158.

Hancock, C., Kaput, J., & Goldsmith, L. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 2*(3), 337-364.

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2004). Data Seen through Different Lenses. Scientific Reasoning Research Institute, University of Massachusetts, Amherst. TERC.

Makar, K., & Confrey, J. (2005, 2-7 July). *Using Distributions as Statistical Evidence in Well-structured and Ill-structured Problems.* Paper presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL), Auckland, New Zealand.

McClain, K. (2005, 2-7 July). *The Evolution of Teachers' Understandings of Statistical Data Analysis: A Focus on Distribution.* Paper presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL), Auckland, New Zealand.

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45*, 103-129.

Pfannkuch, M. (2005, 2-7 July). *Informal Inferential Reasoning: A Case Study*. Paper presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL), Auckland, New Zealand.

Pfannkuch, M., Arnold, P., & Wild, C. (2011). Statistics: It's reasoning not calculating. (Summary research report on Building students' inferential reasoning: Levels 5 and 6) Retrieved from www.tlri.org.nz

Pfannkuch, M., Regan, M., Wild, C., & Horton, N. (2010). Telling data stories: essential dialogues for comparative reasoning. *Journal of Statistics Education, 18*(1). Retrieved from http://www.amstat.org/publications/jse/v18n1/pfannkuch.pdf.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: from a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.

Roth, W.-M. (2005). *Doing qualitative research : praxis of method*. Rotterdam: Sense Publishers.

Rubin, A., Hammerman, J., Puttick, G., & Campbell, C. (2005, 2-7 July). *Developing models of distributions using Tinkerplots.* Paper presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL), Auckland, New Zealand.

Uniservices asTTle team (2008). Quality questioning using the SOLO taxonomy. Retrieved from http://www.slideshare.net/jocelynam/solo-taxonomy

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10-26.