A Critical Step in Students' Reasoning about Distribution: Moving from Understanding to Using for Inference

Chris Reading
The National Centre of Science, Information and Communication Technology and Mathematics Education for Rural and Regional Australia
University of New England, NSW, Australia, 2351
creading@une.edu.au

Jackie Reid
School of Mathematics, Statistics and Computer Science
University of New England, NSW, Australia, 2351
jreid@turing.une.edu.au

1. Introduction

Recently researchers in statistics education have been focusing on how students reason about various statistical concepts, in particular distribution. The *Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy* (SRTL5) focused on reasoning about distribution and the proceedings (Makar, 2005), as well as a special issue (volume 5 number 2) of the *Statistics Education Research Journal* (SERJ) in November 2006, reported the nature of this reasoning and how it might be supported. Included in the reported research was a hierarchy developed from student responses and proposed as a framework for assessing the cognitive development of students' reasoning about distribution (Reading & Reid, 2006). This paper revisits that hierarchy and presents research that aimed to provide information to assist educators to recognise how students progress through the hierarchy.

2. Research background

Before outlining the previously-developed hierarchy on which this study is based, distribution is defined and research into associated reasoning is summarised. First, insights are provided into the meaning of the concept 'distribution'. Typically, tertiary introductory statistics textbooks define the distribution of a variable as "the values that it takes and how often it takes those values" (Moore & McCabe, 2003, p. 5) and then expand the definition for probability distributions by using proportions (p. 306) rather than frequencies. Basic features expected in descriptions of distributions (p. 12) are the overall pattern (i.e., shape, centre and spread) and deviations from the pattern (e.g., outliers). However, when researchers Bakker and Gravemeijer (2004) investigated the concept of distribution, they identified centre, spread, density and skewness, as key elements. These, combined with the typical definition and assuming that density and skewness provide information about shape, provide a framework for the concept of distribution consisting of five key elements: centre, spread, density, skewness and outliers.

Next, some leading research into reasoning about distribution explains how better reasoning can be facilitated. To nurture conceptual understanding of the key elements of distribution, necessary before reasoning can develop, students must be provided with opportunities to reason about distributions in different contexts (Bakker & Gravemeijer, 2004). In particular, Wild (2006) believed that multivariate situations gave more purpose to student investigations about distribution than univariate situations, and Pfannkuch (2006) emphasized the need to move from the less formal to the more formal when building reasoning towards inference. The importance of developing the notion of distribution, and recognizing the interrelation among key elements before students can compare and analyse data sets, were recognized by Leavy (2006, p. 106).

The final focus is on the measurement of the cognitive development of reasoning about distribution. In statistics a student's understanding of a concept is critical to any future cognitive application of that concept. Models of cognitive development for various statistical concepts have commonly identified two cycles of development, one involving development of the understanding of the concept and the second (more cognitively developed) involving the application of that concept. These concepts have included distribution (Reading & Reid, 2006), analysing and interpreting data (Jones, Langrall, Mooney & Thornton, 2004) and data handling (Watson, Collis, Callingham & Moritz, 1995). Each of these models was based on the SOLO Taxonomy (Biggs & Collis, 1991) with a two-cycle application as described by Pegg (2003, p. 245). The

cycles of levels depend on a relational set of links being interconnected between the elements of interest before increased cognitive activity can take place. This interrelatedness of meaningful elements was also recognized by Bude (2006), in relation to statistical concepts, as important in his proposed three-level continuum of understanding, where the move from the second to the third level required a profound understanding of the concept ready for the concept to be used.

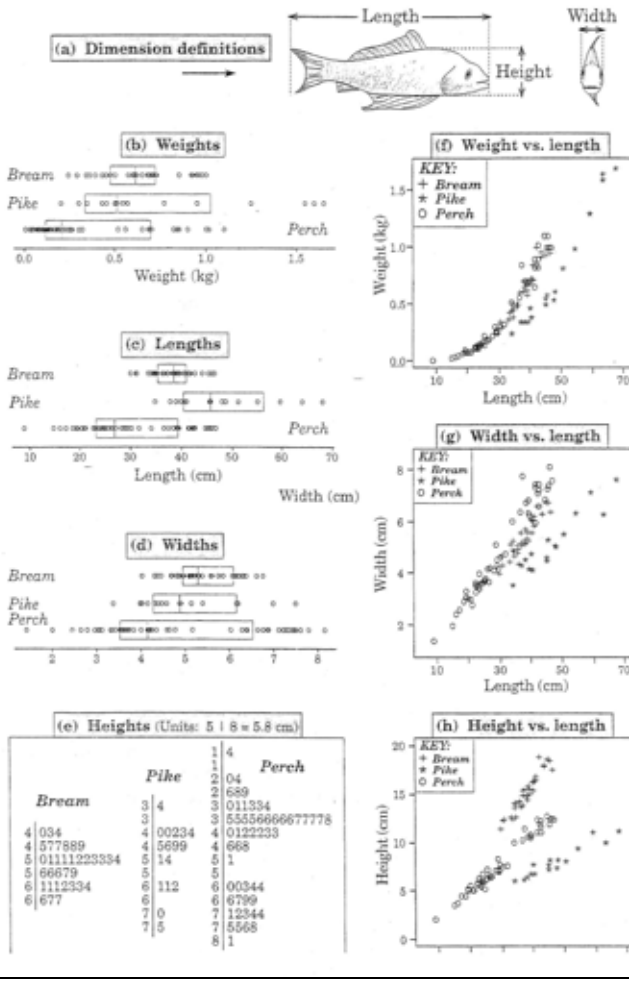
For the concept of distribution Reading and Reid (2006, p. 58-62) described the need for relational linking of the various elements of distribution (end of first cycle) before distribution could be used for statistical inference (second cycle). Their study was based around a regional Australian university one-semester introductory statistics course that was structured around four main themes: exploratory data analysis, probability, sampling distributions, and inferential statistics. Their two-cycle *Hierarchy of Reasoning about Distribution* (figure 1) was developed (Reading & Reid, 2006), using the SOLO Taxonomy framework (Biggs & Collis, 1991), from student responses to minute papers given during each of the main themes. The study involved students from a variety of science-based disciplines who studied in the course, either as compulsory or as an elective. Results of the study were reported in detail in Reading and Reid (2006) and Reading and Reid (2005). The first cycle of developing reasoning, was the based around the understanding of the key elements of distribution and the second cycle was based around using distribution for statistical inference. Before students are able to use distribution for statistical inference (second cycle), they need to have a relational understanding of the key elements of distribution (achieved in the first cycle). If educators want students to be able to operate in the second cycle, i.e., use the concept of distribution when making statistical inferences, then it is crucial that students are able to make the move from the first to the second cycle. Hence, the research question: What does the move from Cycle 1 to Cycle 2 look like in the classroom? In particular, what can student discussion indicate about students' understanding of distribution and about their readiness to use distribution for making inferences?

CYCLE 1	*Understanding the key elements of distribution*
Prestructural (P1)	does not refer to key elements of distribution
Unistructural (U1)	focuses on one key element of distribution (centre, spread, density, skewness or outliers)
Multistructural (M1)	focuses on more than one key element of distribution
Relational (R1)	develops relational links between various key elements of distribution
CYCLE 2	*Using distribution for statistical inference*
Prestructural (P2)	recognizes the concept of distribution but does not use it to make inferential statements
Unistructural (U2)	makes one inferential statement described in such a way as to indicate a correct understanding of the concept of distribution
Multistructural (M2)	makes more than one inferential statement described in such a way as to indicate a correct understanding of the concept of distribution

Figure 1. Hierarchy of Reasoning about Distribution (Reading & Reid, 2006)

3. Methodology

During the same tertiary first year statistics service course as reported in Reading and Reid (2006), tutorial questions were completed by the students for each of the four key themes. In each tutorial the tutor summarized key ideas from the lectures. The tutorial question (figure 2) reported here is the question given during the exploratory data analysis theme and involved students describing key features of both univariate and bivariate plots. For this question, students were only required to describe the main features of the univariate plots (b) and (c), and bivariate plots (f) and (h). To investigate how students responded a case study approach was taken, where two pairs of students were videotaped as they worked together. Of the total enrolment (n=207) in the course about 10% agreed to be considered for videotaping but only two pairs of students were videotaped, due to resource restrictions. Final selection was based on convenience afforded by tutorial scheduling with the two pairs being selected from different tutorial groups. No special protocol was necessary for the passive videotaping of the discussion between each pair of students as they worked on the pre-set in-class tutorial questions. The videotaped sessions were analyzed to determine whether there were indications of reasoning about distribution as described by the hierarchy in figure 1. This analysis was exploratory in that student interactions were being investigated, rather than individual student responses.



The plots give information about the dimensions of three species of fish (bream, pike and perch) caught in a lake in Finland. The data are a subset of the fishcatch dataset submitted by Juha Puranen to the *Journal of Statistics Education* data archive. The variables used in the plots are length, height, width and weight. Working in pairs, describe the main features of the following plots (b) weights, (c) lengths, (f) weight vs length and (h) height vs length.

* diagrams from Wild & Seber (2000, p. 136)

Figure 2. Tutorial Question

4. Results

First, some background is given about the two pairs of students. Then, the main features of the reasoning about distribution found in the students' transcripts are discussed. The expectation was that the students would collaborate with their partner but the reality was that collaboration was limited because these students were not accustomed to discussing statistical ideas with their peers. One pair consisted of two female students, Anna and Naomi (pseudonyms). Both were enrolled in the Bachelor of Science, did well in assignment and examination work and achieved creditable final grades. When it came to answering the tutorial question, Naomi dominated the discussion giving Anna very little opportunity to express her ideas. For example (figure 3), when Anna tried to draw Naomi's attention to two data points that may be extreme values she was cut off by Naomi who identified the lack of whiskers (in the boxplot) as an issue and then moved on.

Naomi:	And these ones go as high as these two on the Pike one up here (*points to plot (d) in error*). They're all by themselves. There's nothing else there
Anna:	Maybe
Naomi:	This one's just bimodal with a really big range. That's the biggest range. (*refers to Perch*)
Anna:	Also, there's two down here
Naomi:	There could be two outliers but depends on where the whiskers went to. Since we don't have the whisker marks…o.k.

Figure 3. Naomi and Anna discussing plot (c) lengths

The other pair consisted of two male students, Ben and Matt (pseudonyms), enrolled in the combined Bachelor of Arts/Bachelor of Science and the Bachelor of Science respectively. Both withdrew before the end of the course. During the tutorial activities the discussion between Ben and Matt was more balanced than for the female pair, with each allowing the other to contribute. This provided a more supportive environment for each to explore his ideas. For example (figure 4), when Matt described the Bream distribution as *even*, Ben agreed by restating that feature as *fairly consistent*. Matt described the same distribution as *fairly compact* and then Ben agreed by restating as *no extreme values*. Matt described the Pike distribution as *sort of more spread out* and Ben agreed with has *a large range*. Both males were struggling for the correct terminology but were able to recognize the important features. This was particularly obvious when Ben refered to *direct concentrations* and *specific concentrations* as he tried to describe the density. During their interactions the males generally used names to refer to the variables while the females tended to point to indicate particular variables.

Ben:	So, I guess you would say for Bream
Matt:	There's an even distribution.
Ben:	Yeah. It's fairly consistent
Matt:	It's fairly compact as well
Ben:	No extreme values
Matt:	Whereas, Pike's sort of more spread out
Ben:	It has a large range. There is really no direct concentrations or anything
Matt:	Nope
Ben:	No specific concentrations
Ben:	I suppose you've got a high concentration down there. You've also got a high sample range compared to the other two I think (*writing*)
Ben:	There are quite a few extreme values in those two
Matt:	Yeah

Figure 4. Ben and Matt discussing the plot (b) weights

Univariate plots

For the univariate plots (see (b) and (c) in figure 2), students' reasoning about distribution included: using various key elements of distribution in the descriptions, linking these elements, making comparative statements and looking for causes to explain features. There was considerable description of the key elements. Centre was not discussed much and was indicated by the standard term *median*, and by non-standard terms such as *all the way down here*. Some reference was made to spread, for example, as *spread* and *range*. Matt's use of the term *compact* is interpreted as a reference to a small range because he reinforced *fairly compact* with *no extreme values* and also linked *range* and *very compact* later on. Far more attention was given to density. The standard terms unimodal and bimodal were used by Naomi, whose observation that *it's got the two sections* indicated that she clearly understood her use of the term bimodal. However, a variety of non-standard references were made to density. The females tended to use *condensed* while Ben used the term *concentration* prolifically. Matt twice referred to *even distribution*, which together with Ben's *fairly random* and *fairly consistent*, were used to describe the spacing of the data and suggest consideration of density. Both Anna and Naomi referred comfortably to skewness on a number of occasions. The males, however, did not use the term at all, although Matt suggested skewness with *a median back here somewhere but its spread all the way out past here*. Outliers were referred to explicitly by Naomi, and alternate expressions such as *extreme value(s)* (Ben) and *all by themselves* (Naomi) were also used.

What indications are there, from the univariate plot discussions, that students have developed an understanding of distribution (first cycle) and are ready to use distribution to make inferences (second cycle)? Sometimes the students attempted to link the key elements, e.g., Matt linked centre (*median*) and spread (*all the way out past here*) to imply skewness. But more often key elements were used in comparative statements. Mostly comparisons clarified relative magnitudes, such as to demonstrate 'how' large, e.g., Matt described the Pike *as more spread out* with reference to the Bream; Ben described a *high sample range* for Pike *compared to the other two*; and Naomi clarified the *lot larger* Pike by comparing them *to the Perch*

which are all the way down here. Less common were comparisons of different features of the plots, e.g., Naomi pointed out that the *largest one* for Perch is *only with the median* for Pike. No formal inferences were made, although Ben tried to explain the cause of an observed feature in the data when he attributed the bimodal distribution of the Perch to the *males and females* (see figure 5).

Ben:	It's fairly random again *(refers to Pike)*
Matt:	And a smaller range
Ben:	Yeah
Ben:	There's still not direct sort of …
Ben:	There's a reasonable concentration … they're similar … like the males and females I guess *(refers to Perch)*
Matt:	mmm
Ben:	Still looks bit like …two separate concentrations I guess

Figure 5. Ben and Matt discussing the plot (c) lengths

Interpretation of some of the examples above was aided by using the context of surrounding comments. For example when determining whether *in the middle* (Naomi) meant that all of the data were in the middle of the distribution or that the distribution was located in the middle of the horizontal axis, her comments about the *median*, and *right skew* suggested the latter, i.e., location of the data. This is reinforced by the fact that both Naomi and Anna then add *condensed* as the next feature, indicating that they had moved on to considering where the data are located within the distribution.

The boxplot representation may not have assisted the males greatly in their interpretations as Matt had to ask how to interpret them and Ben's response that it indicated *where the range area is* was misleading. The females were better able to interpret the boxplots but the reduced format, with no whiskers, appeared to be a drawback when they were trying to determine whether data points were outliers.

Bivariate plots

For the bivariate plots (see (f) and (h) in figure 2), discussion focused on the systematic and random components of the models the students visualised, equivalent to the centre, and spread and density, respectively for the univariate data distribution. There was also some use of these components in comparison and inference, as well as a search for causes to explain the features. When considering the systematic component, the females recognised a trend in plot (f) with Naomi describing it as *exponential*. For plot (h), Anna described the trend as *linear* and Naomi compared the gradients. However, Naomi struggled to articulate the range of the data for the bivariate case. The males indicated trend with *weight increases… length increases* (Ben) and references to gradients (Matt).

There is little indication that the males considered the random component, except perhaps when Matt described the Perch as *fairly uniform, like fairly even*. However, this comment appears to have come after Matt had been struggling with the terminology in relation to the gradients, and may still have been referring just to the different gradients of each data set. The females, however, considered the random component specifically. For example, Naomi stated that *[the data] fit the trend… they're not scattered all over* and referred to how *close to the line* the data were. Anna finally introduced the notion of *random component* and Naomi reinforced this with *random error.* The amount of random error was described as *pretty clustered* (Naomi), with *none out there or out there*. However, Naomi dismissed the possibility of a bimodal distribution (density) even though the Perch data in plot (h) contains two clusters.

What indications are there, from the bivariate plot discussions, that students have developed an understanding of distribution (first cycle) and are ready to use distribution to make inferences (second cycle)? There was little use of the key elements for drawing inferences, although both pairs attempted to make comparisons to explain that there are in fact different relationships. The females used the gradient to explain their inference that the trends were different (see figure 6 and note that Naomi dominated the discussion), while the males used gradient to justify that there were *three distinct groups*. There is also some evidence that the students were looking for causes to explain the distributions; Naomi suggested that the linear trend might have had something to do with *anatomy* while Ben suggested that the different behaviour of Pike in plot (f) may just be the result of small sample size producing extreme values.

Naomi:	o.k. So you've got a linear trend one and..(*writing*)
Naomi:	What's this like down here? The longer they get the higher they get
Naomi:	But I don't know. That will have something to do with their anatomy but
Naomi:	Yeah but how do you describe those ones are higher, that start up higher? How do you say that? Rather than these ones start right down low
Naomi:	The way … the degree to which they increase. How do you …
Naomi:	Because the gradient on this line isn't as great as say this one. The gradient is like…the…the slope.
Naomi:	o.k. So the slopes of the lines are different. Is that a good answer?
Anna:	I don't know
Naomi:	I don't know whether it's a good statistical analysis
Naomi:	The slopes of the lines … (*writing*)… o.k.
Anna:	It looks as though
Naomi:	So the Pike and then the Perch and then the Bream.

Figure 6. Naomi and Anna discussing plot (h) height versus length

5. Discussion

To assist students to move from merely understanding distribution (cycle one) to being able to use distribution when making inferences (cycle two), educators need to be able recognise how students articulate their reasoning about distribution. This exploratory study has described what the move from the first to the second cycle may look like in the classroom, in particular, has described features of student discussion that indicate understanding of distribution and readiness to use distribution for making inferences. Indications from the analysis are that the students had developed a good understanding of distribution but were struggling to relate the various key elements of distribution, an essential cognitive process to progress to being able to use distribution for inference. While some key elements may be referred to using the correct terminology, students may use less formal terminology, or even descriptive phrases, to describe a key element (figure 7) even when the element is understood.

key element	*informal terminology*
centre	median, all the way down here, exponential, linear
spread	spread, range, close to the line
shape	fairly compact, unimodal, bimodal, condensed, concentration, fairly even, fairly uniform, fit the trend, they're not scattered all over, pretty clustered, not out there or out there
skewness	median back here somewhere but its spread all the way out past here
outliers	extreme values, all by themselves

Figure 7. Informal terminology for key elements

More importantly, specific linking of the key elements as indicative of well-developed understanding, an indication that students are ready to make use of distribution for inference, was only sparingly observed in many discussions. However, the use of key elements in comparative statements, an informal form of inference, showed that the students were ready to use key elements of distribution to make inferences. For the univariate data these informal comparisons included the magnitude and location of key elements of distribution. The bivariate plot discussions presented clearer uses of distribution to suggest inferences, such as using the gradient to claim there were three distinct groups, whereas with the univariate data the students were more content to settle for a description of the data. However, none of these discussions were of a sufficient quality to be able to claim that formal inference had been made. Poor articulation makes it difficult to determine whether understanding exists, e.g., the girls ranked the three species of fish (based on gradient) but did not explain what inference they were making about the species based on this comparison. Educators need to be sympathetic of students' immature articulations and provide experiences that allow inferences to be articulated more formally. Similarly, Wild (2006) recommended that multivariate data situations were a richer environment for students when working with distributions.

In summary, educators know that students understand distribution and are ready to use distribution to make inferences when the students can link the key elements of distribution. Although, the students may need assistance with improving their use of the formal terminology required for statistical discussions, they have sufficient understanding of distribution to at least make informal inferences. Learning situations that support the move from informal to formal inference become important, as Pfannkuch (2006) recommended, for students to become proficient suers of distribution to make inferences. Educators should be aware, as they listen to students articulate their understanding, that although a student may confidently direct discussion and hence sound ready to move onto that important step of using distribution to make inferences, a deeper analysis of the content of the discussion may reveal a lack of understanding. This was the case with Naomi who confidently made statements about the key elements but was not able to engage in conversations with Anna when challenged about any of her claims.

This study was exploratory in nature and the results should be viewed in the light of the study's limitations. First, the task analysed did not specifically require students to make inferences. All inferences made by the students were inspired by their attempts at description of the plots and a natural inclination to draw conclusions was more evident for the bivariate plots. Second, the students were not experienced at discussing their ideas with each other. This impacted especially on the pair of girls where one dominated the exchanges and stifled the expression of ideas by the other. While this may have prevented the expression of the optimal level of reasoning by the girls, it was still possible to use the girls' discussion to identify expressions of understanding of distribution. Finally,

6. Implications

Although this study was exploratory it has provided some insight into the way students express their understanding of distribution and begin to use distribution to make inferences. These findings provide a guide to educators for recognizing students' reasoning and provide the stimulus for researchers to further investigate how students' reasoning develops. Educators should use the expressions reported in this study to help them to become more aware of how students discuss distribution. This should help the educators to recognise when the students have a good understanding of distribution and are ready to use of distribution to make inferences. Educators also need to be accepting of the informal terminology that students use to express more formal concepts. To assist students, educators need to take care with the selection of learning experiences. Researchers can assist educators by identifying those learning experiences that allow students to move from the understanding of distribution (cycle one) to the use of distribution to make inferences (cycle two). Researchers should also seek to further identify the expressions that students use, both in their individual responses and in discussions with others, to reason about distribution to further develop the descriptions of the two cycles of reasoning about distribution and, in particular, what is crucial to the step from the first to second cycle. A first step to inform such research would be to explore student discussions later in the course, after the students have completed learning activities designed to reinforce the concept of distribution and introduce the use of distribution when making inferences.

ACKNOWLEDGEMENTS

This research was funded by a Science Faculty Internal Research Grant from the University of New England.

REFERENCES

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.

Biggs, J., & Collis, K. (1991). Multimodal learning and the quality of intelligent behaviour. In H. Rowe (Ed.), *Intelligence, reconceptualization and measurement* (pp. 57-76). New Jersey: Laurence Erlbaum Assoc.

Bude, L. (2006). Assessing students' understanding of statistics. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics: Working cooperatively in statistics education, Salvador, Brazil.* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Jones, G. A., Langrall, C. W., Mooney, E.S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97-117). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Leavy, A. (2006). Using data comparison to support a focus on distribution: Examining preservice teachers' understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal, 5*(2), 89-114.

Makar, K. (Ed.) (2005). *Reasoning about Distribution: A collection of studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4),* Auckland, 2-7 July 2005, [CDROM, with video segments]. Brisbane, Australia: University of Queensland.

Moore, D. S., & McCabe, G. P. (2003). Introduction to the practice of statistics. New York: W.H. Freeman and Company.

Pfannkuch, M. (2006). Comparing boxplot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45.

Pegg, J. (2003). Assessment in mathematics: A developmental approach. In J. Royer (Ed.), *Mathematical Cognition* (pp. 227-259). Greenwich, CT: Information Age Publishing.

Reading, C., & Reid, J. (2005). Reasoning about variation: A key to unlocking the mystery of distributions. In K. Makar (Ed.), *Reasoning about Distribution: A collection of studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4),* Auckland, 2-7 July 2005, [CDROM, with video segments]. Brisbane, Australia: University of Queensland.

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

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation, 1*, 247-275.

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

Wild, C. J., & Seber, G. A. F. (2000). *Chance encounters: A first course in data analysis and inference.* New York: John Wiley & Sons.

ABSTRACT

Distribution is a key concept in statistics. An operational understanding of distribution is critical for students to be able to confidently engage in statistical inference. Researchers are now providing hierarchies to assist statistics educators to assess a student's level of reasoning and hence support the development of a deeper level of reasoning. One such hierarchy of reasoning, about distribution, is the focus in this research. Analysis of discussion by two pairs of students as they worked through a tutorial question for a tertiary-level introductory statistics course, indicated that they were well advanced in an early cycle of reasoning about distribution. The results of this research assists statistics educators to identify when students are ready to move from a basic understanding of the concept of distribution, to being able to use distribution for statistical inference.