CHARACTERIZING YEAR 11 STUDENTS' EVALUATION OF A STATISTICAL PROCESS

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

ABSTRACT

Evaluating the statistical process is considered a higher order skill and has received little emphasis in instruction. This study analyses thirty 15-year-old students' responses to two statistics assessment tasks, which required evaluation of a statistical investigation. The SOLO taxonomy is used as a framework to develop a hierarchy of responses. Focusing on the quality of response allowed insight into and suggestions for how instruction might be improved. The implications for teaching, assessment, and the curriculum are discussed.

Keywords**:** *Statistics education research; Evaluating statistical investigations; Assessment; SOLO taxonomy; Secondary students*

1. INTRODUCTION

In 2002 a new approach to national assessment in New Zealand was introduced at Year 11 (15-year-olds). Instead of one final external examination in mathematics, one third of the course is now internally assessed, with external moderation, and the rest is an external examination (New Zealand Qualifications Authority, 2001). Statistics is internally assessed and students are given data sets to investigate. The assessment is standards-based with three performance levels: achievement, merit, and excellence. Achievement requires students to interpret statistical information and answer straightforward questions. For merit, students must also draw inferences, justify their answer to their question, and comment on features in the data, whereas for excellence the requirement is to evaluate the statistical process. Students must provide evidence that they can meet these levels in two tasks. The tasks are designed so that it is clear what performance level each question is measuring. The level of statistical thinking required at Year 11 with this new internal assessment, compared to the previous external assessment that largely asked students to read and interpret graphs and calculate measures of central tendency, has produced real challenges for teachers and students. The focus of this paper is on characterizing student responses to the excellence part of the assessment, which requires students to evaluate the statistical process.

1.1. RELATED RESEARCH

Evaluation of the statistical process (problem, plan, analysis, conclusion) requires thinking tools such as a list of criteria or "worry questions" for each stage (see Section 4, Wild & Pfannkuch, 1999). These thinking tools need to be an integral part of students' analytic techniques that can be triggered to stimulate thought processes on what issues

Statistics Education Research Journal, 4(2), 5-26, http://www.stat.auckland.ac.nz/serj

need to be considered and taken into account when conducting a statistical investigation. Wild and Pfannkuch (1999) proposed that a checklist of basic questions could be drawn up for students which could be underlain with more and more sophisticated questions in an internet-type procedure. These underlying questions could be accessed as students' understanding progressed. This proposal emerged from the realization that the statistics discipline had developed tools for the analysis stage of the statistical process but had not paid attention to developing analytic tools for the other stages of the investigative cycle. One discipline that has developed such tools is quality management. The students in this study did not have access to any thinking tools for the problem, plan, and conclusion stages and this raises the question about what issues beginning students will consider when they evaluate a statistical investigation.

Gal and Garfield (1997, p. 4) stated that students should "be aware of possible biases or limitations or the generalizations that can be drawn from the data" but according to Gal (1997, p. 49) "little has been written about issues involved in assessing students' opinions about data." Research has been carried out on assessing students' opinions about media articles (e.g., Watson, Collis, & Moritz, 1994) but there is limited research on evaluating and analyzing the quality of students' opinions about a statistical investigation. People have written about the assessment procedures used when students conduct their own statistical investigations (e.g., Starkings, 1997; Holmes, 1997) but have not reported an analysis of students' responses. However, the critical evaluation of statistically-based reports in relation to statistical literacy has been a recent focus in research, and since there is considerable overlap in the skills required between evaluating someone else's report and one's own statistical investigation this literature will be drawn upon.

For the interpretation of media reports Watson (1997) identified a three-tier hierarchy of skills. These skills were: basic understanding of terminology; embedding of language and concepts in a wider context; and questioning claims. The first two skills are relevant for interpretation of the problem by the students but the third skill of challenging claims presented in the media is only partially relevant as the students would be challenging their own claims. From another perspective, Gal (2002, p. 3-4) believes that critical evaluation of statistically-based information is predicated on "a *knowledge* component (comprised of five cognitive elements: literacy skills, statistical knowledge, mathematical knowledge, context knowledge, and critical questions) and a *dispositional* component (comprised of two elements: critical stance, and beliefs and attitudes)."

Considering Gal's (2002) perspective, each of the five cognitive elements of the knowledge component is elaborated upon with respect to how each can be used as a criterion for the setting of assessment tasks. A Year 11 student's capability and the prescribed curriculum are also taken into account. For general literacy, the first cognitive element, students need to understand the text as well as distinguish the meaning of statistical terms (e.g., spread) from their everyday meaning (Watson, 1997). A written assessment task for students should ensure the text is written in a meaningful way for the particular age group and that the statistical terms that are used in the text are part of their statistical knowledge base, the second cognitive element. When evaluating a statistical investigation students may be required to draw upon a wider statistical knowledge base such as having knowledge about sampling variability. The statistical knowledge base element, which Year 11 students are currently exposed to, is problematic. Questions, for example, have been raised about the type of conceptual experiences Year 11 should have as they move towards formal inference (Pfannkuch, 2005). Mathematics knowledge, the third cognitive element, at its basic level refers to 'number sense', which refers to an ability to correctly interpret numbers such as fractions and percentages in a report (Gal, 2002), which would be assumed knowledge at Year 11. At another level 'number sense'

means evaluating whether the data or numbers are plausible, which requires an ability to spot basic arithmetic errors, inconsistencies, and massaging of data, knowledge that probably cannot be assumed.

Context knowledge, the fourth cognitive element, is not only necessary for interpreting and gleaning information from statistical data but also is a prerequisite for critical reflection about statistical information (Wild & Pfannkuch, 1999; Pryor, 2001; Gal, 2002). Therefore the Year 11 assessment tasks should use contexts that are sufficiently well known to students that their 'real world' knowledge could be used not only to understand the problem but also to suggest possible improvements or alternative explanations in their evaluations. Pryor (2001), in her research on tertiary students' ability to critique media reports, identified critical thinking from a context knowledge base as a precursor to critical thinking from a statistical knowledge base. Gal's (2002) list of critical questions, with which students should be familiar, addresses the fifth cognitive element. The list contains mainly statistical knowledge worries but it would be justifiable to have more context knowledge worries. This raises the question as to what critical worry questions would be suitable for Year 11 students when evaluating a statistical process.

When considering Gal's (2002) dispositional component for statistical literacy, Wild and Pfannkuch (1999) claimed that a person's propensities to adopt a critical stance and to be curious and imaginative were dispositions that drive a statistical investigation. Hence, it would seem that for Year 11 students to evaluate an investigation the adoption of a critical stance would help them in their ability to critique an investigation. An implication is that the assessment tasks should be sufficiently motivating and interesting to the students to invoke a critical stance and, if possible, the tasks should also challenge their beliefs. According to Pfannkuch (1996), students' non-awareness of their own beliefs and attitudes or of community assumptions affect their ability to evaluate media reports. Such findings have implications for teaching the evaluation of a statistical process. A willingness to think beyond one's own beliefs at the metacognitive level should be part of students' learning experiences in the classroom. Indeed, Gal (1997) believed that the development of students' ability to generate sensible and justifiable opinions should be a focus of instruction. He suggested that teachers should first elicit the student's opinion and then follow up with a question asking the student to provide evidence for the opinion. A climate of "explaining one's reasoning" should be fostered in the classroom in order for students to learn how to evaluate a statistical process.

Evaluation implies that there exist criteria upon which judgements are made (Bloom, 1956). This raises the question as to what criteria should be used for evaluating the statistical process. Starkings (1997, p. 144) stated in her marking schedule the criterion for evaluation of the statistical process: "Clearly relates solution to the problem. Shows a good understanding and appreciation of the solution." The New Zealand Qualifications Authority (2001) marking schedule exemplars referred to sources of bias, improvements, limitations, and appropriateness of the statistical process and a few suggestions were given on how a student might answer a particular question. These examples were general in that they could be applied to any evaluation such as stating another graph that could be drawn, more accurate measurements that could be taken, or more data that should be collected. The ability to evaluate a statistical process is considered to be indicative of achieving "excellence" in the given task. When considering the exemplars given to teachers, however, the judgement of excellence does not seem to be based on a high level quality of response.

Pegg (2003, p. 252) stated that the SOLO model (Biggs & Collis, 1982) not only offered a method to categorize the quality of the responses but also allowed "teachers an

insight into where instruction might most profitably be directed." It is this twofold applicability that is pertinent to this research. First, teaching evaluation of the statistical process is new to teachers and hence they are uncertain what cognitive level or patterns of thought are present in their students and what constitutes a quality response. Second, such an analysis will aid their understanding of how to foster and scaffold students' thinking in the evaluation of a statistical process. From a research perspective more knowledge will be built up in this area.

1.2. RESEARCH QUESTIONS

As part of a larger project on developing Year 11 students' statistical thinking, the following three research questions are addressed in relation to the evaluation of the statistical process and are based on responses to two assessment questions:

- What response category types describe Year 11 students' evaluation of a statistical process?
- What issues do students consider when evaluating a statistical process?
- What SOLO levels do students attain when evaluating a statistical process?

2. RESEARCH METHOD

The research described in this paper is concerned with an identified problematic area from the first year of a planned three-year project on developing students' statistical thinking. The Wild and Pfannkuch (1999) statistical-thinking framework underpins the research project. The framework is initially employed to communicate to teachers the nature of statistical thinking and habits of thinking that should be fostered in students. It is then not only concretized by the teachers in their instruction but also is employed as a thinking tool to critically reflect upon and to describe and analyze teaching and learning situations.

2.1. APPROACH TO RESEARCH

A developmental research method is used that is based on the ideas of Gravemeijer (1998), Wittmann (1998), and Skovsmose and Borba (2000) (see Pfannkuch & Horring, in press, for a fuller account). The research method is developmental in that an action-research cycle is set up whereby problematic areas are identified by teachers and researcher through observations and critical reflections on the implementation of a teaching unit and by the researcher through analysis of student assessment responses. The students also identify areas of concern about their learning through a questionnaire. The teachers and researcher then discuss how the current situation might be changed for the following year when the unit is taught again. The teachers then rewrite the teaching unit.

An initial approach to the mathematics teachers by the researcher during 2002 resulted in them selecting Year 11 for the project. The case-study teacher was self-selected. A workshop, which focused on communicating the nature of statistical thinking to the teachers, was conducted by the researcher. After the workshop the case-study teacher and another teacher were interviewed to identify problematic areas in their 2002 statistics-teaching unit (Pfannkuch & Wild, 2003). These two teachers and the researcher then discussed teaching ideas that could be implemented to enhance the development of students' statistical thinking. The teachers wrote a new four-week statistics unit for 2003. Although all Year 11 teachers implemented the new teaching unit research data were mainly collected from the case study classroom. These data were videotapes of 15

lessons, student bookwork, student responses to the assessment tasks, student questionnaires, and the teacher's weekly audio-taped reflections on the teaching of the unit. Two main areas of concern, identified by the researcher and case-study teacher after the first teaching implementation in 2003, were informal inference and evaluation of the statistical process. Thus the first analysis of these data focused on these identified problematic areas. The informal inference analysis and its implications are reported in Pfannkuch (2005).

2.2. PARTICIPANTS

The school involved in the project draws on students from low socio-economic backgrounds, is culturally diverse, and has teachers interested in improving their statistics teaching. This secondary girls' school like many other schools in Auckland city has a high percentage of new immigrants to New Zealand (about 60%), many of whom have English as their second or third language. In the case-study classroom there were thirty students who were regarded by their teacher as above average in mathematical ability. In this particular class 45% were Pakeha (New Zealand European), 40% were Maori (New Zealand indigenous) or Pasifika (Pacific Islands), and 15% were Asian or Indian. Two students chose not to be video-taped for the research project. The teacher is Pakeha, in her mid-thirties, has a first degree majoring in education, a Masters degree in mathematics education, and has taught secondary mathematics for twelve years.

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. She also writes the internal assessment tasks, which are moderated at the national level. The researcher previously knew the teacher on a professional basis. The researcher was used as a source of teaching ideas before and during the teaching of the unit and was consulted about the assessment tasks.

2.3. THE ASSESSMENT TASKS

Students were given two assessment tasks, Task One (Appendix A) and Task Two (Appendix B) which were created by the case-study teacher. The assessment occurred in two stages. In the first stage the students were given only the story and data for Part A of Task One and asked to pose a question. The teacher then marked their ability to pose a question. For students who could not pose a question, the teacher gave a question to them. In the second stage the students were given one hour and forty minutes to complete both Part B of Task One and Parts A and B of Task Two.

For Task One the students were given a table of data showing the maximum temperatures of two cities Napier and Wellington, which were taken from some summer newspapers. A story involving a decision about where to go for a summer holiday was communicated to the students. Students were required to pose a question (e.g., Which city has the higher maximum temperatures in summer?), analyze the data, draw a conclusion, justify the conclusion with three supporting statements, and evaluate the statistical process with three statements (see Appendix A, Task One, Question 4). For Task One it should be noted that the data were presented to the students as two independent samples. A statistician might have recorded the maximum temperatures for each city each day and then conducted a paired comparison test.

Task Two had two parts. In Part A of the task weather data from the Pacific, Australia, and New Zealand regions were used to generate questions for the students to

answer not given in Appendix B). In Part B students were required to evaluate the statistical process carried out by another person, named Jason (see Appendix B, Task Two, Question 2). For the evaluation of the statistical process it was decided to prompt the students to consider each stage of the process (problem, plan, analysis, conclusion). The prompt for considering the problem posed was omitted from the second task but on reflection should have been included. This research is focused on the students' evaluation of a statistical process and hence it is the student responses to Question 4 of Task One and Question 2 of Task Two that are analyzed.

2.4. APPROACH TO ANALYSIS OF ASSESSMENT RESPONSES

The analysis of the evaluation of the statistical process occurred in two stages. First, the student assessment responses to the two evaluation questions were analyzed. The analysis used a spreadsheet whereby a clustering procedure was used to sort the responses into categories (Miles & Huberman, 1994). The classification of the quality of the response for each level within a category used a hierarchical performance level approach based on the SOLO taxonomy (Biggs & Collis, 1982). The approach recognizes that within the concrete-symbolic mode, in which these students would most likely be functioning, there are at least one and possibly two distinguishable cycles of thinking operating through four levels (PUMR): pre-structural (P) – no use of relevant aspects; unistructural (U) – focuses on one piece of relevant data; multistructural (M) – two or more pieces of data used without integration; relational (R) – all data integrated into coherent whole (Pegg, 2003). These hierarchical levels were determined again by using a clustering approach within the spreadsheet. Based on the student responses qualitative descriptors for each level within a category were written and coded by the author, and then another person independently coded all responses. A consensus was reached between them on the final codes for each student response.

Second, the transcriptions of the video-tape data from the case-study classroom were qualitatively analyzed for instances of the evaluation process in operation in the classroom. This analysis was used to inform the discussion about the assessment responses.

3. RESULTS

The three research questions are addressed respectively in this section. First, descriptors of the category types for student responses to the evaluation of a statistical process, which were derived from the data, are discussed. Second, examples of the student responses are discussed in terms of the issues students considered when evaluating the statistical process, and third, a summary of the SOLO levels attained by the students is presented.

3.1. RESPONSE CATEGORY TYPES

Within the four stages of the statistical process most responses were to the analysis and plan stages, giving four distinct categories whereas there was little response to the problem and conclusion stages, which were combined into one category. The five main categories of response identified with respect to the students' evaluation of the statistical process were: *My/Someone Else's Analysis* and *Another Analysis* that could be conducted

for the analysis stage; *More Data* and *Other Data* that could be collected for the plan stage; and *Other* which mainly related to the problem and conclusion stages (Fig. 1). These categories turned out to be similar to the prompts given to the students in the questions. Within these categories hierarchies of responses were identified and qualitatively described, reflecting the use, combining, and relating of elements suggested in the SOLO model. The descriptors for all the categories were similar in that they followed a sequence of *specify*, *justify*, and *relate*. The latter three categories, however, relied mainly on contextual knowledge of the situation whereas the former relied mainly on statistical knowledge. A possible transition into a higher-level mode requiring statistical knowledge, which was more abstract than contextual knowledge, was also identified for the latter three categories (Fig. 1).

	Evaluation of the Statistical Process			
	Analysis Stage Categories		**Plan Stage Categories**	**Problem/ Conclusion Stage Category**
SOLO Level Description	**1. My /Someone Else's Analysis**	**2. Another Analysis**	**3. More Data 4. Other Data**	**5. Other**
Prestructural (P) Inappropriate response	Gives an inappropriate reason why analysis is a good/bad choice.	Gives an inappropriate improvement or non-specific improvement.	Gives an inappropriate improvement or non-specific improvement.	
Unistructural (U) Single elements	Specifies one appropriate reason why analysis is a good/bad choice.	Specifies one appropriate statistical improvement.	Specifies one appropriate contextual improvement.	
Multistructural (M) Multiple elements	Justifies/critiques the choice of analysis in relation to the original question.	Justifies or gives an appropriate reason for the statistical improvement.	Justifies or gives an appropriate contextual reason for the improvement or an appropriate broad statistical justification.	
Relational (R) Relates to investigation	Justifies/critiques the choice of analysis in relation to the information that can be derived from that analysis or to the ability to reason from that analysis to answer the original question.	Relates the improvement to the original question under consideration.	Relates the improvement to the original question under consideration.	
Extended Abstract (U(2)) Brings in extra statistical elements			Specifies a statistical improvement.	

Figure 1. Categories and hierarchical descriptors for evaluation of the statistical process

The hierarchical descriptors for each category will now be explicated more fully in terms of the student responses. The data suggested it was necessary to have a separate descriptor for *My/Someone Else's Analysis* as full integration at the relational level seemed to occur when the students amplified how a particular analysis allowed them to reason about the question, such as this response for Task One:

> S5: I believe that the box-and-whisker graph was the most appropriate graph to use because it is very easy to read and at a glance you can see that Napier is overall warmer than Wellington. It is an appropriate graph for a comparison question (R).

The *Another Analysis* hierarchical descriptors were similar to categories 3, 4, and 5. The difference was that the students were commenting on the analysis and hence needed to use their statistical knowledge to justify the suggested improvement. An example of each level of response for Task One is:

> S21: Use histogram graph (P).
>
> S17: A back to back stem-and-leaf may have been a better graph because it would have shown all the figures (U).
>
> S18: Could draw a stem-and-leaf graph (back to back) and look to see if there are any peaks (M).
>
> S29: A back to back stem-and–leaf would also have been a good graph for me to draw because it would have shown me the shape of the data and given me a good idea where most of the temperatures were for each city (i.e. 20 something degrees or something teen degrees etc.) (R).

The prestructural response was considered an inappropriate improvement, as it did not clearly specify how the data would be compared. The unistructural response gave an appropriate alternative graph for the comparison of data but the reason was inappropriate. The multistructural response recognized that a stem-and leaf graph allows peaks to be seen implying that these could not be seen in boxplot graphs. The relational response extended the idea further by relating this advantage to being able to find out more information about the temperatures of the cities.

For categories 3, 4, and 5 one set of descriptor levels was sufficient. The first identified cycle was based on and characterized mainly by contextual knowledge of the situation. Occasionally a student gave a broad statistical justification in the sense that the statement could apply generally to any investigation and hence it was classified as multistructural within the first cycle rather than a second cycle response. The following response was classified as multistructural in the first cycle since such a general statement was considered a broad statistical justification rather than a specific and full statistical justification and explanation:

> S2: His analysis can be improved if he had more information of temperatures from other days and maybe also from another country so that he will be able to generalize his findings (M).

In comparison, the beginnings of specific statistical improvements were identified in two students' responses. These students seemed to be moving beyond the relational level as they began to think about how they might analyze those data, for example, suggesting the possibility of a graph:

S24: As my friend and I are wanting to know the warmest place to go perhaps it would be to our benefit to collect a range of data and graph the coldest temperatures of these two cities as well. That way we would find out how cold it may get, and this may well alter our perspective of where we wanted to travel on holiday. That is other data that could also be collected (U(2)).

Such a response was classified as U(2) because it went beyond the level expected in the concrete-symbolic mode by bringing in the beginnings of statistical knowledge rather than being solely based on contextual knowledge.

3.2. ISSUES CONSIDERED BY STUDENTS

Considering the students were evaluating the statistical process (problem, plan, analysis, conclusion) their responses to and the criteria they used for judging each stage of the process are highlighted. It should be noted that *students were only required to make three statements for each question*. Therefore the no response category means that a student did not respond in that particular category.

Problem Stage Improvements to the question posed were classified with the conclusions stage under *Other* as few students responded to this stage. Two students responded successfully by suggesting an improvement to the question they posed for Task One (see Table 1), for example:

S23: Next time I would improve the question I posed by looking into the maximum temperatures not trying to draw conclusions on finding a warmer climate with only maximum temperature statistics. I would change my question to, does Napier or Wellington have a higher maximum temperature? Since that is more to do with the data I was given (U(2)).

The two successful students realized that the measures used were possibly not relevant to the question they had posed.

Table 1. Task One: Details of student responses

	Analysis Stage		Plan Stage		Problem/ Conclusion Stages
SOLO Level	My Analysis	Another Analysis	More Data	Other Data	Other
No response	8	22	10	12	23
Prestructural	2	1	13	4	5
Unistructural	4	5	3	6	
Multistructural	13	1	3	5	
Relational	3	1	1	2	1
Unistructural(2)				1	1
Total number of students	30	30	30	30	30

Plan Stage Improvements to the plan centered on whether *More Data* or *Other Data* should be collected before making a decision or drawing a conclusion. In specifying *More Data* that should be collected, the student responses revealed a prevalent

misconception. Twelve students for Task One and two students for Task Two (Table 2) mentioned that the sample size should be the same for each data set. A typical response for Task One was:

S6: In the data given there were 3 temperatures not given for Wellington. If they were given, the statistics could have increased or decreased and affected the results. There should have been an even amount of data for both sides – Wellington and Napier (P).

Table 2. Task Two: Details of student responses

	Analysis Stage		Plan Stage		Problem/ Conclusion Stages
SOLO Level	Someone else's Analysis	Another Analysis	More Data	Other Data	Other
No response	13	8	18	24	27
Prestructural	3	6	3	3	1
Unistructural	7	6	6	2	2
Multistructural	7	9	2	1	
Relational Unistructural(2)		1	1		
Total number of students	30	30	30	30	30

The more successful students focused on whether a reasonable sampling method had been used and suggested what data should be collected and why:

S2: I think that the analysis can be improved if she had another set of data to compare because temperatures can vary anytime of the year (M).

This response was considered a borderline multistructural response as the student did not clearly state that temperatures should be collected from other years and could vary from year to year. Acknowledging that temperatures could vary, however, is the beginnings of understanding sampling variation from a contextual perspective.

For the category *Other Data,* over half the students responded in Task One and one-fifth of the students responded in Task Two. Some students did not specify the actual data that should be collected and hence their responses were classified as prestructural such as the following statement for Task Two:

S4: Other data could have been collected to verify or support Jason's statement in a more trustworthy way (P).

If the student was able to specify the appropriate weather data to collect the response was considered to be unistructural:

S22: Jason could have improved this by: Using data from the whole world not just NZ and Pacific / Oz (U).

This response suggested an emergent realization that proving a theory in one region of the world was insufficient and that such an observation should be replicated elsewhere. Specifying appropriate weather data to collect and giving a reason that related to the warmth of the climate, such as the following response to Task One, was considered to be multistructural:

> S26: Other data that could be collected to improve the analysis is rainfall over summer because to me places can be humid and raining, it would be important to know other aspects of the weather to compare regions (M).

Such responses from students indicated that they were beginning to realize that capturing the notion of 'warm' with a single measure was insufficient and that other measures for warmth such as humidity and minimum temperatures, should be considered in the comparison of regions. A relational response was considered to be a coherent whole when the specified data and reason for collecting them were justified in terms of the question posed:

> S5: It may be that Napier gets colder during the nights than Wellington does. The minimum temperature should also have been gathered as the posed question asked "Which city is warmer over the summer period, Napier or Wellington?" and this data does not give the adequate information to correctly answer the question (R).

Eighteen student responses were classified under *Other Data* for Task One. Of those students, eight were classified at multistructural and above (Table 1).

Analysis Stage Giving reasons why *My Analysis* was a good choice in Task One (22 students) and suggesting *Another Analysis* for Task Two (22 students) prompted the most response (Tables 1 & 2). Most responses for these categories focused on suggesting either that the student's own graph was the best choice for Task One or a box-and-whisker graph was more appropriate for Task Two. A typical multistructural response for *My Analysis* in Task One was:

> S7: I think I made the best choice in picking a box-and-whisker graph as it clearly shows the comparison between Napier and Wellington and their temperatures (M).

In Task Two in the *Another Analysis* category a prevalent multistructural response was:

> S15: He should have drawn a box-and-whisker so you could actually compare the results (M).

The notion that box-and-whisker plots were useful for comparing grouped data was a typical response, with 16 and 10 students responding in Task One and Task Two respectively at the multistructural level and above.
For *Someone Else's Analysis* in Task Two, however, there were four comments on the categorization of the data such as:

> S30: Could have kept the shower/rain in a different table and surveyed more days to see if rain affected the outcome (M).

These four students were beginning to realize that the categorization of data was relevant to a statistical analysis and that a different categorization might produce a different conclusion.

Conclusion Stage. Responses to the conclusion classified under *Other* were limited. One student attempted an alternative explanation for Jason's theory in Task Two, which was very convoluted, but showed she was willing to challenge the assumption that the clouds keep in the heat. Part of her response was:

> S5: My point is that it may not actually be the clouds keeping in the heat – it may just be that when it's colder it's less likely to heat up or get dramatically colder than it is, and that when it's hotter it gets much colder during the night (U).

Another student wondered whether her conclusion for Task One made sense with what she knew about the real world situation and reasoned from an individual event explanation:

> S4: Although these statistics have shown that Napier tends to have a higher maximum temperature than Wellington this statement in my opinion is not all that accurate. Because weather is unpredictable, Wellington might have a nice sunny day but then wind comes along and the temperature drops giving a low reading, e.g. 16.0 rather than 19.9 (P).

Students did not attempt to evaluate whether their conclusions were valid from the perspective of inference space judgement. No student responses were classified at the multistructural level or above for judging the conclusion.

3.3. SUMMARY OF THE SOLO LEVELS ATTAINED BY STUDENTS

The results tables (Tables 1 & 2) give an overview of the level of the responses that students demonstrated according to this method of analysis. It should be noted that students were only required to make six statements with respect to evaluating the statistical process, that is three statements for Task One and three statements for Task Two. In Task One the average was 2.5 statements per student whereas for Task Two it was 2 statements per student. The case-study teacher did observe that some students ran out of time to fully answer the Task Two questions. When considering the data overall there seemed to be some students operating fairly consistently at the same levels in the PUMR cycle. A summary table, calculated by taking the best four statements that a student made out of a possible six statements and assigning 0 to a prestructural or no response, 1 to a unistructural response etc., and then finding the mean score, produced a student profile of the class (Table 3).

Table 3. Summary of student responses

SOLO level and Mean Score	Number of students
Prestructural (0-)	7
Unistructural (0.75 -)	14
Multistructural (1.75 -)	8
Relational (2.75 -)	1
Total	30

Two-thirds of the class appeared to be operating at the unistructural level and below. It should be noted that, even though the case-study teacher's marking schedule for the evaluation questions was not based on these SOLO criteria, only one student was awarded an "excellence" grade, and this was the same student who was categorized as thinking at the relational level.

4. DISCUSSION

The student patterns of thought observed in the data could be said to be in response to the method of instruction (1), or to the students' general or contextual knowledge (2), disposition (3), statistical knowledge (4), cognitive development (5), or general literacy levels (6), which include both text comprehension and ability to communicate, or to the task which specifically mentioned aspects to comment on (7). It was considered that all seven factors could be operating on the level of student response, the contributory effect of each being unknown.

For the first factor, an analysis of the videotape data that recorded the teaching of the unit revealed that the instruction only once briefly focused on the evaluation of the whole statistical process. In one lesson the students compared the prices of second-hand exercycles and home gyms, which were data gathered from the newspaper advertisement columns. The teacher prompted an evaluation of the statistical process by asking: "How could I improve my investigation?" The students suggested these ideas: "Find out what kind of gym"; "Look at it everyday for a couple of weeks"; "Brand"; "Its quality"; and "How old it is". The teacher elaborated on their ideas but at no stage asked them to justify their opinions. In another two lessons the students evaluated graphs. The instruction was:

> Okay, I'm going to give you 5 minutes to look at the graph and I want you to write anything that pops into your mind, okay about the graph. Any conclusions that you can make, anything that you think is confusing on the graph, and also other questions that you might ask. So I want you to think quite generally. I want you to think about conclusions and other things that arise. In fact to start you off, you might want to go back to that phrase "I notice", and "I wonder".

Basically the students focused on whether graphs were misleading, on interpreting the information, and on thinking of contextual reasons for the distribution of the data.

Even though statistical investigations with an evaluation of the investigation have been an internally assessed component at Year 13 for about twenty years, this teaching approach would suggest that the evaluation part has not been a focus of the taught curriculum. Indeed the school textbooks have only cursorily covered this aspect. The main emphasis in textbooks and teaching has been misleading graphs. The current exemplars for Year 11 and the teacher workshops, provided by the New Zealand Qualifications Authority, gave minimal direction to teachers. Hence the outcomes of this study suggest that new understandings of how one teaches evaluation of the statistical process are needed if students are to improve their responses.

The responses to the evaluation task revealed a misconception that was not evident during teaching. Even though students had dealt with data sets of unequal sample size in class, they were never asked to evaluate those investigations and hence the prevalent misconception that data sets should have the same sample size before being compared was not uncovered. When reflecting on this misconception, and thinking of Curcio's (1987) hierarchical model for interpreting graphs, the author's observation, corroborated by the case-study teacher, was that the students had experience of reading the data, less

experience at reading between the data, and little experience of reading beyond the data. If these students had some experience of inferring "missing data" from a data set they may have predicted that the missing summer temperatures were likely to be within the interquartile range or at least within the range. The problems of missing data are well known in statistics and students could be given opportunities to impute values for observations and to analyze data with and without the imputations. Specific attention could be drawn to students' beliefs and to whether their conclusions would change with unequal sample sizes. Although evaluation of the statistical process might be considered a higher level skill (Bloom, 1956) and may not currently be a strong feature of teaching, it would seem that allowing students to express their opinions on the overall investigation might allow some different insights into their thinking. These insights need to be reflected upon critically to determine new teaching approaches.

The other six factors that were identified as possibly affecting student responses raised three main issues. Firstly, cognitive development and the disposition or the willingness to adopt a critical stance might have had an effect on student responses but such effects could only be ascertained through a large longitudinal study. Secondly, the tasks were written with the students' contextual and statistical knowledge and text comprehension in mind but presumably these had an effect on student outcomes. Thirdly, general literacy was observed in the students' ability to communicate. The two coders learned a salutary lesson when they were reaching a consensus about the level assigned to two students. One of the coders was challenged on her ascribed levels for the two students and asked to justify the levels in terms of the given hierarchical descriptors. The coder then realized that the student who did not express herself fluently actually should have been awarded a higher level for her response, and that the student, fluent in English, was awarded one level too high for her response. Thus the level descriptors appeared to ensure that a marker was objective and not swayed by a student's ability to communicate. Pryor (2001) in her research on tertiary students ability to evaluate media reports calculated that text comprehension had twice the impact compared to graph comprehension and critical thinking in predicting students' ability to think statistically. If text comprehension and the ability to communicate the evaluation of the statistical process are related then it could be conjectured that such ability might have some effect on the level of response.

The analysis of the student responses produced a general hierarchy for evaluation of the statistical process described briefly by *specify*, *justify*, and *relate*. Gal (1997) referred to students justifying their opinions so that assessors could judge their reasonableness. This research confirmed his viewpoint in that a multistructural response was considered to be one in which the proposed improvement was justified. A relational response, however, extended the idea further by asking students to relate their opinions to the question under consideration. A similar scheme of argumentation for justifying inferences from data was also proposed by Cobb (1999). Furthermore, the analysis led to the conjecture that evaluating the statistical process for the problem, plan, and conclusion stages of an investigation might, firstly, be built on contextual knowledge, and secondly, on statistical knowledge. It was conjectured that this might be either a second UMR cycle in the concrete-symbolic mode or the beginning of the next mode, the formal mode, in the SOLO model (Biggs & Collis, 1982). Whatever the mode the next level is in, it is hypothesized that the multistructural level may occur when the student is able to give a statistical justification for the improvement mentioned. The relational level might occur when there is full integration of the contextual and statistical justification or critique when related to the original question. Presumably at this level the student would also demonstrate a fluent use of statistical language and ideas. The conjectured integration of

these two knowledge bases at the relational level is supported by Wild and Pfannkuch (1999) who identified it as one of their five fundamental statistical thinking elements. It would seem that for learners contextual knowledge would be more prominent at first, a facet also found by other researchers (e.g., Watson, Collis, Callingham, & Moritz, 1995).

When evaluating media reports, Gal (2002) suggested students should have a list of critical questions in their heads while Wild and Pfannkuch (1999) suggested that students should have thinking tools at their disposal for all stages of the investigative cycle. From the student responses and the task prompts the critical questions for evaluation of the statistical process are shown in Figure 2. These questions were derived from two tasks and might be limited but they seem to be a suitable subset of critical questions at this stage for the Year 11 students.

All seven factors considered above may have affected the level and type of student response. At this stage, the teaching factor is the one area that can be targeted for improvement in the second year of the project. The analysis of the student responses has allowed some insight into their patterns of thought and the hierarchical descriptors have provided a possible structure for teachers to foster students' ability to evaluate the statistical process.

- Could improvements be made to the question? Are the measures used relevant to the question posed?
- Could improvements be made to the method of data collection? Has a reasonable sampling method been used?
- What other data should be considered or collected before making a decision?
- Are there better graphs that could be drawn or other statistics that could be calculated? (If you believe that you have made the best choice of graph(s) and statistics, explain why.)
- Could improvements be made to the categorization of the data?
- Is the conclusion valid? Does my conclusion make sense with what I know about the real world? What are some possible alternative explanations?
- Has the conclusion been drawn about the sample data under consideration?

Figure 2. Proposed judgement criteria for the evaluation of the stages of the statistical process

5. CONCLUSION

From a teaching perspective this analysis with the resultant hierarchical descriptors enabled the writing of model solutions to the evaluation questions of both tasks, which will be used by the teachers. The hierarchical descriptors have explicitly revealed the type of responses sought and will direct their teaching to scaffold students to higher levels of thinking. From an assessment perspective the hierarchical descriptors will enable teachers to be sure that a high quality response is awarded 'excellence' and they will explicitly know why a response is high quality. From a curriculum perspective the analysis raised some questions about the learning experiences and conceptual development at Year 11 for the evaluation of the statistical process. Statisticians, educators, and researchers need to work together on developing a teaching pathway that gradually builds up more critical questions or judgement criteria for evaluating a statistical process, which is directly linked to the prescribed curriculum content. From the statistics discipline perspective the current approach to evaluation is largely unstructured and is reliant on a statistician's

experience in the field. Statisticians should begin to develop thinking tools for the problem, plan, and conclusion stages of the investigative cycle for the general statistics discipline not only for the enhancement of problem-solving but also for the evaluation of their own and others' investigations (Wild & Pfannkuch, 1999). These tools will require a synthesis of contextual and statistical understanding.

Gal (2002) claimed that critical evaluation of media reports was predicated on the joint activation of a knowledge component and a dispositional component. This research suggests that critical evaluation of the statistical process may be predicated on such joint activation but is enacted through communication and evaluative skills. These skills need to be explicitly taught and fostered in instruction with specific attention paid to justifying opinions and relating such justifications to the question under consideration. For the problem, plan, and conclusion stages of a statistical investigation instruction needs to focus on scaffolding students thinking to consider not only contextual but also statistical justifications and specifically explaining those justifications. Such thinking will present real challenges for teaching and the curriculum.

Wild and Pfannkuch (1999) suggested there were four dimensions in statistical thinking: the investigative cycle, types of thinking, the interrogative cycle, and dispositions. The interrogative cycle can be thought of as operating at the micro and macro level whereby the thinker evaluates the statistical process by: generating possibilities, seeking or recalling information, interpreting and connecting ideas, criticizing ideas against contextual knowledge, statistical knowledge, beliefs and so forth, and judging what to believe currently. Further research is needed on eliciting, understanding, and developing students' evaluative thinking. This research is based on a small sample and must be regarded as exploratory. Evaluative thinking, however, is a crucial dimension in fostering students' statistical thinking and deserves more research attention.

ACKNOWLEDGEMENT

The author thanks Jane Watson and Julia Horring for their comments and contributions to this paper.

REFERENCES

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.

Bloom, B. (1956). *Taxonomy of Educational Objectives*. New York: David McKay Company.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, *1*(1), 5-43.

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, *18*(5), 382-393.

Gal, I. (1997). Assessing students' interpretation of data. In B. Phillips (Ed.), *IASE Papers on Statistical Education ICME-8*, Spain, 1996, (pp. 49 - 57). Hawthorn, Australia: Swinburne Press.

Gal, I. (2002). Adults' statistical literacy: meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1-51.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-13). Amsterdam, The Netherlands: IOS Press.

Gravemeijer, K. (1998). Developmental research as a research method. In A. Sierpinska & J. Kilpatrick (Eds.), *Mathematics education as a research domain: A search for identity* (pp. 277-295). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Holmes, P. (1997). Assessing project work by external examiners. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 153-164). Amsterdam, The Netherlands: IOS Press.

Miles, M., & Huberman, M. (1994). *Qualitative Data Analysis*. Thousand Oaks, CA: Sage Publications.

New Zealand Qualifications Authority (2001). *Level 1 achievement standards: Mathematics*. [Online: http://www.nzqa.govt.nz/ncea/ach/mathematics/index.shtml]

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

Pfannkuch, M. (1996). Statistical interpretation of media reports. In J. Neyland & M. Clark (Eds.), *Research in the learning of statistics: Proceedings of the 47th Annual New Zealand Statistical Association Conference* (pp. 67-76). Wellington, New Zealand: Victoria University.

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., & Horring, J. (in press). Developing statistical thinking in a secondary school: A collaborative curriculum development. In G. Burrill (Ed.), *Curricular development in statistics education: Proceedings of the 2004 International Association for Statistical Education Round Table Conference, Lund University, Sweden, 28 June-3 July, 2004*.

Pfannkuch, M., & Wild, C. J., (2003). Statistical thinking: How can we develop it? In *Bulletin of the International Statistical Institute 54th Session Proceedings* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Pryor, H. (2001). *Assessment of the statistical literacy ability of some tertiary students using media reports.* Unpublished Masters Thesis, The University of Auckland.

Skovsmose, O., & Borba, M. (2000). *Research methodology and critical mathematics education.* Publication No. 17 Roskilde, Denmark: Centre for Research in Learning Mathematics, Roskilde University.

Starkings, S. (1997). Assessing student projects. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 139-152). Amsterdam, The Netherlands: IOS Press.

Watson, J. M. (1997). Assessing statistical thinking using the media. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107-121). Amsterdam, The Netherlands: IOS Press.

Watson, J. M., Collis, K., & Moritz, J. (1994). *Authentic assessment in statistics using the media*. Report prepared for the National Center for Research in Mathematical Sciences Education – Models of Authentic Assessment Working Group (University of Wisconsin). Hobart, Australia: University of Tasmania, School of Education.

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

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

Wittmann, E. (1998). Mathematics education as a 'design science'. In A. Sierpinska & J. Kilpatrick (Eds.), *Mathematics education as a research domain: A search for identity* (pp. 87-103). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

APPENDIX A

Task One: Holiday Temperatures **Name: _____**
Maths Teacher _____

Part A

Every year you and a friend argue about where to go on your January summer holiday. You both enjoy outside activities and really enjoy the beach. Next year you will either go to Napier or Wellington for your holiday.

Since you both want to go to the warmest place you decide to analyse the maximum temperatures in Napier and Wellington. Your friend has found a stack of last summer's newspapers. She has gone through them and recorded the maximum temperatures in both places. These are shown in the tables below. Note that the temperatures are not in order.

What *statistical* question or hypothesis could you answer using this data?

Maximum Temp Napier ^{o}C
25.2
24.5
22.0
24.5
21.7
22.8
22.9
24.6
24.1
25.2
23.8
20.2
23.9
19.9
23.6
25.8
21.2
22.7
23.4
28.7
21.4
27.6
22.8
22.8
22.9
23.0
26.4
25.8
27.3
20.5
28.9
29.6
33.1

Maximum Temp Wellington ^{o}C
21.6
21.5
20.9
22.0
23.5
18.8
18.0
22.2
19.2
24.0
24.6
19.5
24.6
25.0
22.2
21.6
20.5
21.4
19.9
16.1
18.6
19.7
16.0
20.2
21.8
25.6
25.5
27.4
23.6
23.1
Not given
Not given
Not given

Task One: Holiday Temperatures **Name:_____**

Part B

Analyse the data in order to answer your question.

Use the data for Napier and Wellington to *answer your question* or *test your hypothesis*. The following instructions will help you do this.

1. Calculate statistics for Napier and Wellington. These must include
   at least one measure of central tendency
   at least one measure of spread.

2. On the graph paper provided draw appropriate graphs(s) that allow you to answer the question or test the hypothesis you posed.

3. *Respond to your question or hypothesis.* Refer to your statistics and the features of the graph(s). Use these to support your conclusion. Make 3 statements that justify your conclusion.

4. Write an evaluation of the statistical process. Aim to make 3 statements. If you make more than 3 statements *select* the 3 statements which you think are the best.

Your statements could refer to some of the following aspects:
- Other data that could be collected to improve the analysis.
- Improvements to the method of data collection.
- Better graphs that could be drawn, or other statistics that could be calculated. (If you believe that you have made the best choice of graph(s) and statistics explain why).
- The validity of your conclusions.
- Improvements to the question posed.

APPENDIX B

Task Two: Cloud Blanket

Part B

Jason read a European magazine about how clouds act as a warm blanket. The article said that the clouds keep the heat in and therefore prevented the temperature dropping very much. He decided to analyse the data he collected from the newspaper to see if this was true.

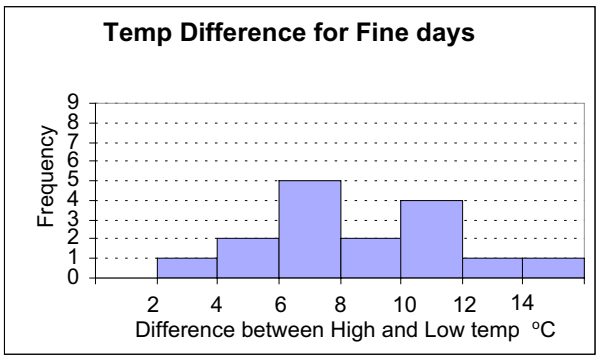
He thought that if the ***clouds keep the heat in*** then the ***difference*** between the high and low temperatures will be ***less*** on cloudy days compared to fine days.

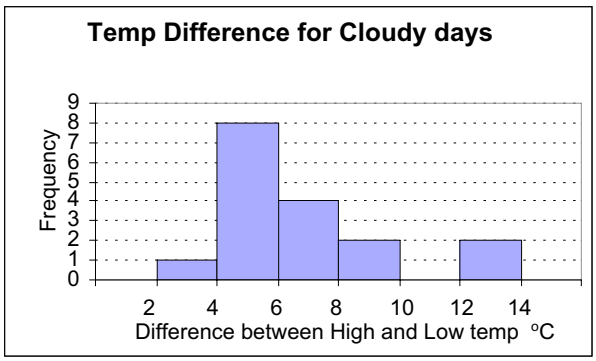
Jason took all the data for the Pacific/Australia and New Zealand and categorised them into 'Fine' and 'Cloudy'. He then took the high and subtracted the low to give the difference. His tables are shown below.

	Clear Sky Temp		
	High	Low	Difference
clear	23	11	12
fine	32	23	9
fine	31	24	7
fine	20	15	5
fine	26	18	8
fine	22	15	7
fine	29	24	5
fine	28	24	4
fine	25	13	12
clearing	13	1	12
fine	18	11	7
fine	19	12	7
fine	17	7	10
fine	18	7	11
fine	18	4	14
fine	16	1	15

	Cloudy/Showers Temp		
	High	Low	Difference
rain	16	7	9
showers	16	11	5
showers	17	12	5
showers	17	11	6
showers	13	4	9
showers	11	6	5
cloudy	16	10	6
showers	31	25	6
showers	27	20	7
showers	14	10	4
showers	19	13	6
showers	31	23	8
showers	30	22	8
showers	31	24	7
showers	29	24	5
cloudy	25	11	14
cloudy	18	5	13

Jason drew the following graphs with this data. He also calculated a few statistics.





Fine Days	°C
Mean difference	9.1
Median difference	8.5

Cloudy Days	°C
Mean difference	7.2
Median difference	6.0

1. Comment on Jason's theory that the clouds keep the heat in. Justify all comments using features of the graphs and/or statistics. (Make 3 statements)

2. Write an evaluation of the statistical process that Jason used for his theory that the clouds keep the heat in. Aim to make 3 statements about how his analysis can be improved. If you make more than 3 *select* the best 3 statements.

Your statements could refer to some of the following aspects:
- Other data that could be collected to improve the analysis.
- Improvements to the method of data collection.
- Better graphs that could be drawn, or other statistics that could be calculated. (If you believe that he has made the best choice of graph(s) and statistics explain why)
- The validity of your conclusions.