MEASURING THE DEVELOPMENT OF STUDENTS' CONSIDERATION OF VARIATION

JACKIE REID
School of Science and Technology, University of New England, Australia
jreid@turing.une.edu.au

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

ABSTRACT

Research investigating how students begin to consider and reason about variation will help educators identify stages of this development. This can provide direction for learning activities to help students develop a strong consideration of variation that can be applied in a variety of contexts. In the present study, tertiary student responses to a class test and an assignment question are analysed, resulting in a description of levels of consideration of variation relevant to these tasks. This and other hierarchies previously developed are used to formulate a Consideration of Variation Hierarchy *applicable to a variety of tasks. Implications for research and teaching are discussed.*

Keywords: *Statistics education research; Consideration of Variation Hierarchy; Statistics education; Tertiary education*

1. INTRODUCTION

An important issue in statistics education, and related research, is how to help students develop statistical thinking, reasoning and literacy. Literature in this area is extensive (e.g., Garfield & Ben-Zvi, 2004; Chance, 2002; Garfield, 2002; Rumsey, 2002). The importance of variation was flagged when consideration of variation was proposed as one of the fundamental types of statistical thinking (Wild & Pfannkuch, 1999). Also understanding of variation has been reported as contributing to the development of students' statistical thinking (e.g., Meletiou-Mavrotheris & Lee, 2002; Reading & Reid, 2005; Reading & Shaugnessy, 2004; Torok & Watson, 2000). Many researchers have reinforced this view. Most importantly, variation is taken to be a foundation concept for statistics. Statistics has been described as the "science of variation" (e.g., MacGillivray, 2004) and Bakker (2003) explained that students who did not expect variability would lack "intuition of why one would take a sample or look at a distribution." Finally, there is increasing interest in describing and measuring the development of understanding of variation and an interest in finding ways to help students use their intuitive notions of variability to move towards a more sophisticated notion of reasoning about variation (Garfield, delMas, & Chance, 2007; Reading & Shaughnessy, 2004). Challenging questions for researchers and educators such as: "What does correct reasoning about

© International Association for Statistical Education (IASE/ISI), May, 2008

variability look like? What are ways to assess understanding of variability? . . . What are useful methodologies for studying the understanding of variability?" were posed by Ben-Zvi and Garfield (2004a, p. 4) for the Reasoning about Variation focus at the Third International Research Forum on Statistical, Reasoning and Thinking, reported in the Forum *Proceedings* (Lee, 2004) and two special issues of this journal (Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005).

This paper addresses fundamental questions concerning students' reasoning about variation by developing a hierarchy of consideration of variation. In earlier work (Reading & Reid, 2005; Reid & Reading, 2004, 2006), hierarchies of levels of consideration of variation were developed, based on tertiary students' responses to minute papers and a questionnaire. In the present study, student responses to a class test and an assignment question are analysed, resulting in a description of levels of consideration of variation relevant to those tasks. This and the other hierarchies previously developed are used to formulate a *Consideration of Variation Hierarchy* applicable to a variety of tasks. Implications for research and teaching are discussed.

2. RESEARCH BACKGROUND

The following provides a review of current research into the development of students' consideration of variation at the tertiary level and in particular, focuses on recently proposed hierarchies that assess and investigate this development.

2.1. CONSIDERATION OF VARIATION IN THE TEACHING AND LEARNING OF STATISTICS

Much of the research to date on the role of variation in statistical reasoning in education has been at the pre-tertiary level. This research has expressed concern that educators have placed too little emphasis on the notion of variation (e.g., Meletiou-Mavrotheris & Lee, 2002; Torok & Watson, 2000). For example, measures of location have been emphasized to the detriment of consideration of variability (Reading & Shaughnessy, 2004), and there is the potential for the deterministic approach of the mathematics curriculum to have a negative impact on statistics instruction (Meletiou-Mavrotheris & Lee, 2002). Lack of stochastic awareness may leave students embarking on their tertiary statistics education ill-prepared to consider the more advanced notions of the statistical model as a combination of both systematic and random effects. Lack of an appreciation of the complete statistical model will contribute to students viewing statistics as a list of techniques to be learned in isolation (Reading & Reid, 2005). A sound understanding of variation could help promote a more comprehensive approach to learning statistics. The four components of Wild and Pfannkuch's (1999) consideration of variation provide a suitable basis for expanding on the notion of understanding of variation. These components are:

1. noticing and acknowledging variation – recognizing the omnipresence of variation and the need to record this variation in discussions;
2. measuring and modeling variation for the purposes of prediction, explanation, or control – creating summaries (numerical or graphical) to represent the variation in the data and using these summaries to represent the impact of variation;
3. explaining and dealing with variation – looking for the causes of variation and considering the impact on design and sampling; and
4. using investigative strategies in relation to variation – formal procedures for looking at the properties of the variation itself.

A thorough assessment of as many as possible of these components of *consideration of variation* should help clarify the development of students' understanding of variation. This approach was taken by Torok and Watson (2000) when developing their categories of the appreciation of variation, and by Reading and Reid (2005) when developing a hierarchy of levels of *consideration of variation*.

Historically, there has been little research that explores the development of students' understanding of variation at the tertiary level. More recently, delMas and Liu (2003) focused their research on tertiary students' interpretations of the standard deviation, and Lann & Falk (2003) found that when students in a first year service course were explicitly asked to consider variation, although their intuitive notions varied, a greater proportion of students chose the range than any other single measure of spread to summarise the variability in a data set. In a broader study of college students' consideration of variation Meletiou-Mavrotheris and Lee (2002) found that students took a more deterministic approach to exploratory data analysis but, although students struggled with concepts of variation in most contexts, by the end of the course, many had an increased awareness of the need for information regarding the spread of a distribution.

Recent trends indicate the use of less traditional strategies for both teaching and assessment in statistics (Garfield & Gal, 1999). Importantly, assessment should be aligned with learning goals, and then the type of instruction and activities required to achieve these goals should be chosen (Garfield & Ben-Zvi, 2004). New tools are required to assess deeper understandings being articulated in these goals. For example, interviews are valuable to gain a better idea of students' understanding (Reading & Reid, 2006a). Information on deeper understandings of variation, such as those based on statistical reasoning and thinking, is crucial for the development and refinement of new curriculum and assessment approaches. It is important to use a range of assessment tasks to examine students' understanding of variation because "… assessment of thinking about variation is heavily reliant upon both the types of assessment tasks employed and the context in which the tasks are situated" (Meletiou-Mavrotheris & Lee, 2002, p. 33). Furthermore, a variety of assessment tasks addressed in different settings would allow educators to better determine further development of instruction and assessment (Reading & Shaugnessy, 2004).

2.2. ASSESSING TERTIARY STUDENTS' CONSIDERATION OF VARIATION

The research project, *Understanding of Variation*, explored the development of tertiary students' consideration of variation as they engaged in the various learning activities and assessment tasks in an introductory service statistics course with 'consideration of variation' as a core for the curriculum. The project aimed to develop and refine hierarchies being developed to assess students' understanding of variation and to investigate this understanding. The project included analysis of student responses to a range of tasks; pre-study and post-study questionnaires, follow-up interviews of selected students, four separate minute papers, one question from a class test, and one question from an assignment. It is important for students to be able to understand and apply the concept of variation in a variety of contexts. The tasks were not designed specifically to focus on variation but rather they were tasks that formed part of the course assessment. The researchers looked for any consideration of variation, that is, the expressions of variation and how these were used, in students' written or verbal responses.

Details are now provided of two hierarchies that evolved from student responses to minute papers and a questionnaire, respectively. Reading and Reid (2005) described levels of consideration of variation (Table 1) based on responses given to the minute

papers (short answer questions given in class). The minute paper questions reflected the curriculum themes of exploratory data analysis (minute paper 1 – MP1), probability (MP2), sampling distributions (MP3) and inferential statistics (MP4). Similarly, levels of consideration of variation (Table 2) based on responses given to a pre- and post-study questionnaire were developed (Reid & Reading, 2006). The four question pre-study and post-study questionnaires were identical and focused on variability (Q1), comparing data sets (Q2), sampling (Q3 & Q4) and probability (Q4). Q1 asked for the meaning of variability. Q2 asked for the description and comparison of the timetable performance of two buses with a graphical summary supplied. Q3 asked for an opinion on a statement about observed outcomes of a particular event given demographic information about the population in New Zealand. Q4, with three parts, asked students to make, and justify, predictions about sampling from a mixture of coloured lollies. In both instances the analysis described levels of *no*, *weak*, *developing* and *strong* consideration of variation.

Table 1. Levels of Consideration of Variation (Minute Papers) – Reading & Reid (2005)

No consideration of variation	
MP1&4	discuss the means only as evidence of the inference, with no mention of variation
MP2	do not mention the relevant factors to explain variation of trial outcomes
MP3	do not mention variation in relation to the distribution
Weak consideration of variation	
MP1&4	discuss the amount of variation but don't explain how this justifies the inference
MP2	incorrectly apply relevant factors to explain variation of trial outcomes
MP3	some description of variation that implies how variation influences distribution
Developing consideration of variation	
MP1&4	discuss the amount of variation and explain how this justifies the inference made
MP2	interpret some factors correctly to better explain variation of trial outcomes
MP3	indicate appreciation of variation as representing distribution of values
Strong consideration of variation	
MP1&4	indicate an appreciation of the link between variation and hypothesis testing
MP2	interpret all factors correctly to give good explanation of variation of trial outcomes
MP3	recognize effect of variation on the distribution and relevant factors

In the following section, we describe the current study that produced the levels of consideration of variation based on student responses to the class tests and assignment questions. The information from this study is then combined with descriptions of levels based on responses to minute papers (Table 1) and pre- and post-study questionnaires (Table 2) to develop a hierarchy that can be used to describe the students' developing consideration of variation across a range of tasks.

3. THE STUDY: METHODOLOGY

The research targeted a one-semester introductory service statistics course (enrolment 46) studied by students in science-related fields at a regional Australian university. The course included a variety of topics with four organizing themes: exploratory data analysis, probability, sampling distributions, and inferential statistics. The presentation of the content in the text for the course (Wild & Seber, 2000) was considered to support the course approach, and throughout each topic the lecturer frequently referred to the core concept of variation. Although all enrolled students were expected to complete the various learning activities and assessment tasks as an integral part of the course, responses were only analysed for those students who agreed to 'participate.' Data collection and analysis were performed by the two authors, one of whom was the lecturer in the course.

Table 2. Levels of Consideration of Variation (Questionnaire) – Reid & Reading (2006)

No consideration of variation	
Q1	do not consider any sources of variation
Q2	may refer to a measure of centre, but not to any measure of spread
Q3	do not acknowledge any variation about the expected values
Q4	do not acknowledge any variation about the theoretical or expected outcomes
Weak consideration of variation	
Q1	discuss one source of variation but expression is poor
Q2	refer to the range and/ or basic description of shape
Q3	acknowledge variation and expectations are articulated but not based on given data; look for extraneous causes of variation
Q4	allow for variation but amount suggested is low or high; causes given are extraneous
Developing consideration of variation	
Q1	describe clearly one source of variation (within-group, between-group, controlling factors, measurement error)
Q2	refer to measure of location and more detailed description of spread
Q3	consider variation between expected and observed values and/or identify need for a larger sample or more information
Q4	provide a realistic amount of variation, but may not be centred correctly; reasoning may be based on frequencies rather than proportions
Strong consideration of variation	
Q1	describe clearly more than one source of variation
Q2	provide further information about the distribution, such as explicit proportions
Q3	[not described because no response coded at this level]
Q4	provide a realistic amount of variation, and proportional reasoning is correctly used

This report focuses on the analysis of responses to a class test question and an assignment question that led to the development of descriptions for the levels of consideration of variation, presented in the next section. Both questions were selected for analysis because they had the greatest potential to allow students to provide information about their consideration of variation. The class test question (Appendix A) used in this study required students to describe and compare distributions and was one of three questions in the test. The test was given during the fourth week of a 12 week course, at the end of a topic on exploratory data analysis. Thirty-three students completed the test. Prior to the test, student tutorial experiences included examining a large data set and interpreting graphs such as histograms, dotplots, scatterplots, and stem and leaf plots. The content of lectures also included discussion on the shape of a distribution (symmetric, skewed, bimodal) and the influence of outliers. As class tests were taken at different times, two versions with different data sets (lampshells and caesarean sections) were used to avoid the issues of prior knowledge of the question. The part of the question, common to each version, requiring a response is reproduced in Table 3. Only responses to part (a), describing the shape of the distribution, and part (c), comparing the distributions, were

Table 3. Class Test Question

(a) Describe the shape of the distributions ….
(b) Give the appropriate numerical summary for each distribution. Justify your choice.
(c) Compare the two distributions.
(d) Using the IQR, identify any potential outliers for the distributions …. Show your calculations.

analysed as they were most relevant to the focus on variation. Makar and Confrey (2005) state that distribution gives "a visual representation of the data's variation" (p. 28). Although shape is only one aspect of describing a distribution, students often include a discussion of the variation in the data when asked to describe the shape of the distribution. Consequently, an analysis of student responses to part (a) of the class test could be expected to provide useful information about students' consideration of variation.

Fifteen students completed an assignment with two questions. The assignment question (Appendix B) selected for analysis was based on a one-way analysis of variance, whereas the other was based on a simple linear regression. Both of these topics had been covered in some depth as part of the curriculum. The assignment was submitted at the end of the course, by which time the students had covered all course content, including one-way analysis of variance. Like the class test question, there were two versions of the assignment question: one pertaining to reading programs; and the other pertaining to cuckoo eggs. The part of the question, common to each version, requiring a response is reproduced in Table 4. No word limit was set but there was an emphasis on clearly describing what was shown by the output, including graphics. Students were asked to produce a graphical representation of the data that allowed a comparison of the groups. Part (a) was chosen for analysis.

Table 4. Assignment Question

(a) Summarise the data in a table giving sample sizes, means, and standard deviations. Give an appropriate graphical summary that allows a comparison of the groups. (b) State and check the assumptions of the ANOVA model: i) by constructing normal probability plots for each group. ii) using Bartlett's test. (c) Give appropriate null and alternative hypotheses to compare the different groups (in words and using statistical notation). (d) Run the ANOVA, producing i) a normal probability plot of the residuals. ii) Tukey's pairwise comparisons. (e) With reference to the output from (a) and (d), write a non-technical summary of your conclusions.

Researchers looked for evidence of consideration of variation in students' responses to the two tasks. Initially responses to a particular question were identified as showing no or some consideration of variation. Those responses showing some consideration of variation were then ranked as displaying *weak*, *developing* or *strong* consideration. The common understandings displayed in these responses at a particular level were then used to describe that level of consideration of variation. Once the levels had been described the responses were coded according to these levels. This procedure was based on that used for the minute papers (Reading & Reid, 2005) and the questionnaire (Reid & Reading, 2006). When the researchers disagreed about the coding level of a response they each explained what aspect of the response had caused them to choose a particular level. The ensuing discussion about the interpretation of the response resolved the disagreement in every case.

4. THE STUDY: RESULTS AND DISCUSSION

4.1. RESPONSE CODING

Analysis of the responses to the class test question showed that there was a variety of features of within-group and between-group variation given. Because comparisons of distributions and one-way analysis of variance are both core topics in the curriculum it is not unreasonable to expect some students to be able to describe and use the concepts of within-group and between-group variation both informally and formally. It was not necessary for students to refer to these terms explicitly but rather be able to describe them, and/or refer to their features and ultimately link them.

When referring to the within-group variation some features identified were extremes, outliers, range, skewness, large distribution, majority between certain limits, spread, and symmetry. When referring to the between-group variation some features identified by students were differences between medians, between averages, and between means. The descriptors resulting from the coding of the class test responses and the assignment question responses were similar. This was not unexpected because both questions required students to compare distributions. Those responses that demonstrated some consideration of variation were coded as either *weak* or *developing* (Table 5). No response was coded as *strong*.

Table 5. Levels of Consideration of Variation (Class Test and Assignment Questions)

No consideration of variation general statements which do not display any meaningful consideration of variation
Weak consideration of variation identify features of either within-group variation or between-group variation; expression used may be poor; terms used may be incorrect or confused
Developing consideration of variation discuss both within-group variation and between-group variation without linking them; refer to variation to support inference but do not link within-group and between-group variation
Strong consideration of variation [not described because no response coded at this level]

As the class tests were completed during non-compulsory class time not all students completed them and consequently only thirty-three responses were analysed. There were only fifteen assignment question responses analysed because, although most students produced the required numerical and graphical summaries, many did not make the comparison, which was the focus of the coding of responses. On the assumption that this might have been a misinterpretation of the question (the wording *"allows a comparison"* may have been ambiguous) these nil responses were not coded at all rather than coding them as *no consideration of variation*. The majority of responses (more than 95%) show some evidence of consideration of variation; however, no response demonstrated what could be considered a *strong* consideration of variation.

Following are examples of *weak* and *developing* responses for part (a) and part (c) of the class test, and for the assignment question. Examples have been selected to demonstrate what might be expected of responses at each level. Each response has an identification tag that begins with R, and then an identification code. The identification code for a test question response is followed by "a" or "c" to indicate whether it was a response to part (a) or (c) of the question, and the data set used is indicated by

(Caesarean) or (Lampshells). Identification codes not followed by (a) or (c) refer to assignment question responses and are labeled (Reading) or (Cuckoo) depending on the data set used. For example, R2a (Caesarean) indicates a response to part (a) of the class test question that used the Caesarean data set, whereas R4 (Cuckoo) refers to a response to the assignment question that used the Cuckoo data set.

4.2. WEAK RESPONSES

Typically, responses showing *weak* consideration of variation presented features of only one of within-group variation or between-group variation. Usually this was within-group variation and the features used to describe the within-group variation depended on the shape of the data. For example, R2a noted the existence of outliers in the distribution of caesareans performed by male doctors, whereas R3c compared the amount of clustering evident in the two distributions. A typical response to the assignment question was R4, which grouped all data from the 6 groups into a single stem and leaf display, resulting in all data being considered as one sample. This representation prevented any identification of between-group variation, and consequently, only discussion of within-group variation was possible. Those *weak* responses presenting features of between-group variation usually compared measures of location. For example, R5a compared the average number of caesareans.

R2a (Caesarean) *For male doctors the distribution is positively skewed with two observations that could possibly be outliers. For female doctors the distribution is roughly symmetrical with a slight positive skew.*

R3c (Caesarean) *Female distribution is highly clustered therefore less variability male distribution is less clustered which shows high variability. More males data was collected. The data shows that more male doctors perform caesarean sections on the whole.*

R4 (Cuckoo) *The decimal point is at the |*

19|69
20|113
20|699999
21|11111134
21|666699999999
22|11111111111111111133333333334444444
22|6666699999
23|111111111113333333344444
23|66999999
24|111111134
24|9
25|1

The above stem and leaf display shows that the lengths are nearly symmetrical, with the majority of egg lengths between 21 and 23 mm.

R5a (Caesarean) *The shape of both distributions is such that there is only one distinct peak in each. This indication that for the majority of both males and females the average number of caesareans performed is similar*.

Some *weak* responses were transitional to *developing* consideration of variation. As well as one of within-group or between-group variation being identified, there was some indication that the other was also being considered. For example, R6a discussed shape and

also demonstrated that means and medians have been considered, but not effectively compared, suggesting that the between-group variation may have been considered.

R6a (Caesarean) *The shape of the distribution for male doctors is bimodal with two peaks, and also a gap between the two peaks, the distribution for males is not symmetrical. The shape of the distribution for female doctors is also much closer to being symmetrical (mean is almost equal to median) than that of the distribution for male doctors.*

4.3. DEVELOPING RESPONSES

The *developing* responses presented features of both within-group variation and between-group variation. Typically these responses gave some description of the variation in each sample and also compared some measure of location for the distributions. For example, R7a mentioned the spread over the whole range for the live lampshells and R8a compared the values above which 50% of the data lie. Typical was response R9c that compared the ranges and the means for the two distributions. Less common was response R10c that considered the within-group variation in terms of how *proportions* of observations are arranged around the average. R11 provided separate consideration of features of both within-group variation (skewed and outliers) and between-group variation (medians centred around middle of boxplots, 'sizes' are smaller). However, there was no attempt to link the two to provide a more detailed comparison of the groups.

R7a (Lampshells) *The live lampshells have quite a bit of variability, bi-modal. They are spread out over the whole range and also have a much larger SD than the dead ones. The dead lampshells are more unimodal with a couple of possible outliers the SD is much smaller and there is not as much variability.*

R8a (Lampshells) *Live lampshells have a bimodal distribution, this bimodal distribution would be different to dead lampshells because there was more data collected on live than dead. The dead lampshells have a negatively skewed distribution with 50% of its data above 20, where live lampshells has 50% of its data above 14.74.*

R9c (Lampshells) *Due to shorter range in dead lampshell and a Large mean, they die at a longer length. However the live Lampshells has a larger range and the mean is smaller then the dead. Therefore lampshell will grow without dying at a young age.*

R10c (Caesarean) *On average male doctors performed more caesarean sections than female doctors. In terms of proportions the female doctors had less deviation around the average than the males did.*

R11 (Cuckoo Eggs) *From figure 1.1 we can see all the different species are roughly normally distributed, with medians centred around the middle of the boxplots. The other groups are slightly skewed with the Meadow Pipit and Hedge Sparrow both recording outliers. We can also see that the sizes of the Cuckoo eggs in the Wrens nest are smaller than the other five species.*

A few *developing* responses were identified as transitional to *strong* consideration of variation because they brought together position as well as within-group variation, indicating an awareness of the need to link within-group variation and between-group variation although they did not do so. For example, R12c discussed variability within each of the two groups, as well as overlap of the distributions, while indicating a comparison of the two to obtain an informal conclusion.

R12c (Lampshells) *The live lampshells have a greater variability than that of the dead lampshells. It cannot clearly be said that dead lampshells are larger than live ones as there is too much overlap in the data. It can be seen that you will find smaller live lampshells than dead ones, probably because they will usually reach a reasonable age and length before they die. The smaller live lampshells are most likely the younger ones*.

4.4. DISCUSSION

Students need to develop a sound consideration of variation and be able to apply it in a variety of contexts. The proposed levels of consideration of variation (Table 5) arose from coding responses to two assessment tasks, according to the consideration of variation exhibited. This analysis has shown that different levels of consideration of variation exist and that these levels represent cognitive development of the concept. The progression from weaker to stronger consideration of variation is marked by improved use of terminology, reference to more than one type of variation, recognition of the need for taking variation into account when making inferences, and the linking of different forms of variation.

Both part (c) of the class test and part (a) of the assignment question required students to compare two distributions, a precursor to a more formal analysis of variance. Although no response was coded as strong, a *strong* response would be expected to link within-group and between-group variation, moving towards an intuitive analysis of variance. The wording of the tasks had an impact on the quality of responses and has implications for the results.

In their responses to part (a) of the class test, many students provided a single-word descriptor for the shape of the distribution (e.g., skewed, symmetric, bimodal). Given the wording of the question ("Describe the shape of the distributions") it is not surprising that many students did not discuss variation in any detail. Furthermore, because part (c) of the same question asked for a comparison, it is unlikely that students would elaborate on links between within-group and between-group variation in part (a). In other words, the form of part (a) of the class test question did not encourage students to demonstrate a more developed consideration of variation. This was also true of the assignment question. Students may not have felt it necessary to include a comparison of within-group and between-group variation in their responses to part (a) of the assignment question because part (d) asked for a formal test (analysis of variance) to compare the groups. The impact of question structure on the amount of information about consideration of variation that student responses can exhibit has also been discussed in Reid and Reading (2004).

5. REFINING LEVEL DESCRIPTORS TO FORM A HIERARCHY OF CONSIDERATION OF VARIATION

This paper has presented research that explored the development of tertiary students' consideration of variation as they engaged in various learning activities and assessment tasks. First, the levels of consideration of variation that evolved from the analysis, in earlier studies, of student responses to minute papers (Table 1) and a questionnaire (Table 2) were presented. Next, the evolution of level descriptors based on responses to class test and assignment questions that asked students to compare distributions (Table 5) was described. These three descriptions of levels evolved from different tasks set in a variety of contexts. Previously, a hierarchy applicable across tasks was proposed by Reid and Reading (2005). Now, more detailed analysis of responses to the class test and assignment

questions, as well as closer interrogation of the levels developed from other tasks, has allowed refinement of the level descriptors resulting in a combined hierarchy. The level descriptors of this hierarchy are justified in the following section.

5.1. JUSTIFYING THE LEVEL DESCRIPTORS

To define each level of this combined hierarchy, the descriptors for the corresponding level in each of the three earlier hierarchies (Tables 1, 2 & 5) were compared. In the following, the elements common to the three hierarchies (Tables 1, 2 & 5) are described, and differences highlighted for each level. In light of this comparison, the process of refinement of the level descriptors is then discussed.

No consideration of variation. All descriptors for this level were very similar, in that responses failed to acknowledge any variation.

Weak consideration of variation. All responses coded at this level, regardless of the task, acknowledged the existence of variation but discussion was generally limited to a basic description of variation (e.g., range), or the description was incorrectly or poorly expressed. These responses indicated awareness that variation exists but suggested a lack of the language and tools necessary to be able to describe or use variation appropriately. It is acknowledged that some 'incorrect' descriptions may be due to lack of expertise with the English language rather than weak consideration. Those educators responsible for students who have English as a second language should take care when coding responses and also interview students to affirm the assessed level of consideration.

Developing consideration of variation. At this level, responses to all tasks provided a more detailed and accurate description of at least one of the two sources of variation, that is, within-group and between-group variation. This was recognized as a minimum requirement for a response to be coded as exhibiting a *developing* consideration of variation in Tables 1 and 2. However, both the class test and the assignment questions required a comparison of distributions. Consequently, in the analysis of the responses to these tasks, it was deemed necessary for a response to include clear references to both within-group and between-group variation for a response to be coded as *developing* (Table 5).

Strong consideration of variation. At this level, the differences among the descriptors used for the various tasks were more pronounced. Using the descriptors presented in Tables 1 and 2, if responses clearly referred to more than one source of variation they would be coded as *strong* responses. There was no descriptor that evolved from the responses to the class test and assignment questions (Table 5) because no responses were coded higher than *developing*. It was anticipated, however, that a *strong* response would link within-group and between-group variation, moving towards an intuitive analysis of variance.

Refining the level descriptors. The preceding comparison makes clear the elements common to the level descriptors across the three hierarchies, but also highlights a number of differences. It was these differences that necessitated a refinement of the descriptors resulting in the *Consideration of Variation Hierarchy* (Table 6). For examples of responses to a variety of tasks at different levels refer to section 4 in this paper, Reading and Reid (2005), and Reid and Reading (2006).

Table 6. Consideration of Variation Hierarchy (combined across all tasks)

No consideration of variation
do not display any meaningful consideration of variation in context
do not acknowledge variation in relation to other concepts (e.g., distribution)
Weak consideration of variation
identify features of only one source of variation (within-group or between-group)
acknowledge variation in relation to other concepts
incorrectly describe variation
do not base description of variation on the data
anticipate unreasonable amount of variation
poorly express description of variation
refer to irrelevant factors to explain variation
incorrectly refer to relevant factors to explain variation
do not use variation to support inference
Developing consideration of variation
clearly describe both within-group and between-group variation
recognize the effect of a change in variation in relation to other concepts
correctly describe variation
base description of variation on the data
anticipate reasonable amount of variation
clearly express description of variation
correctly refer to relevant factors to explain variation
use variation to support inference
do not link the within-group and between-group variation
Strong consideration of variation
link within-group and between-group variation to support inference

In the earlier hierarchies, some of the descriptors refer to aspects of particular tasks. All descriptors that were developed from responses to MP2 (Table 1) refer to trial outcomes (of a coin tossing experiment). Similarly, explicit reference is made to expected and observed outcomes (of births) in the descriptor for *developing* consideration of variation that evolved from responses to Q3 of the questionnaire (Table 2). Furthermore, reference is made to particular statistical concepts: distributional reasoning (Table 1, MP3) and proportional reasoning (Table 2, Q4). For those specific tasks, explicit reference in the descriptors to particular concepts and contexts did not limit the applicability of the descriptors. However, these are not included in the descriptors in Table 6. The more general descriptors in Table 6 still allow the coding of responses that make specific reference to the context of a particular task, or to a particular statistical concept, but also permit the *Consideration of Variation Hierarchy* to be applied to a broader range of tasks.

The first descriptor listed under each of the three levels (*weak, developing and strong)* in Table 6, is considered the key indicator of attainment of that particular level of consideration of variation. So a response is coded as *weak* if it identifies only one of within-group or between-group variation, but as *developing* if it clearly describes both sources of variation. Finally, as part of the development of the *Consideration of Variation Hierarchy*, it was decided that the key descriptor of a *strong* consideration of variation was to be able to *link* within-group and between-group variation to *support* inference. The other descriptors listed under each level provide supporting evidence that a student's response should be coded at that particular level. Not all descriptors for a particular level would necessarily be exhibited in a single response.

The enhancement of the *strong* descriptor, to include the linking of within-group and between-group variation to support inference, reflects a change in the researchers'

expectation of a *strong* response because of a more detailed analysis and comparison of previous level descriptors given in Tables 1, 2 and 5. This has implications when responses to the various tasks used in this study are re-examined using the *Consideration of Variation Hierarchy*. For example, when using the final hierarchy to code responses to Q1 of the questionnaire ("What does variability mean to you? Give a verbal explanation and/or an example."), students would not necessarily be expected to provide a *strong* response, even if they were capable of working at that level, because the question does not require students to make inference. The inclusion of both within-group and between-group variation in the descriptors for *developing* and *strong* consideration of variation does not preclude the use of the hierarchy for coding responses to tasks that consider only a single distribution. However, responses to such tasks could not be coded at the higher levels because the tasks do not require consideration of both within-group and between-group variation. For students to be able to demonstrate the depth of their consideration of variation they need to be provided with tasks that allow them to make inferences involving two or more distributions. Furthermore, it is important to realize that to develop a true picture of a student's level of consideration of variation, responses to a variety of tasks should be considered.

5.2. LIMITATIONS OF THE CONSIDERATION OF VARIATION HIERARCHY

When interpreting and applying the *Consideration of Variation Hierarchy* (Table 6) the limitations of this study, in relation to the level of the statistics course from which it developed and the type of inference tasks implemented, should be taken into consideration. The hierarchy evolved from tasks based on content for an introductory statistics course. There were few responses coded as *strong* and thus there was limited information on which to base descriptors to provide supporting evidence that a response should be coded at the *strong*, rather than the *developing*, level. The hierarchy needs extending to be effectively applicable to tasks from more advanced statistics classes.

Just as the descriptors in the *Consideration of Variation Hierarchy* evolved from earlier descriptors when responses to a greater variety of tasks were analysed, further refinement of the descriptors may be required as student responses to more advanced statistical tasks are analysed. For example, what descriptors are needed to code responses to more complex tasks such as linear mixed models, where students need to take into account both fixed and random factors and consider variance components? Furthermore, all of the tasks required only an informal approach to inference. If tasks requiring a more formal approach to inference were used then the descriptors may need to be further refined to produce a hierarchy that is applicable to an even wider variety of tasks and contexts.

6. IMPLICATIONS FOR TEACHING AND RESEARCH

Consideration of variation is fundamental to the ability to reason statistically. Consequently, teaching and learning activities and assessment items should be structured to address this. This paper has presented a *Consideration of Variation Hierarchy* that evolved from student responses to a variety of tasks, presented in an introductory course that had variation as a core for the curriculum. Educators can use the hierarchy to identify a student's developmental level of consideration of variation. An awareness of the level of development of consideration of variation at which a student is operating can then inform the design and implementation of teaching and learning activities to further that development.

It is not realistic, however, to expect students at the end of a one semester introductory course to consistently exhibit a *strong* consideration of variation. Case studies showed that some students gave no evidence of an improved consideration of variation at the end of the course, whereas others' responses showed an improvement for some, but not all, tasks. Progression in the hierarchy was not a linear process nor was it the same for each student (Reid & Reading, 2005). As Pfannkuch (1997) stated, "…the concept of variation would be subject to development over a long period of time with different tools and different contexts." A challenge for researchers is to investigate hindrances that prevent students from developing a stronger consideration of variation.

The hierarchy was developed from in-class tasks that formed a part of the curriculum. Some of those tasks proved more useful than others in eliciting information about students' consideration of variation. Nonetheless, it is apparent that it is possible to investigate the development of students' consideration of variation, and other statistical concepts (see, for example, Reading & Reid, 2006b), without the need to devise special assessment tasks additional to those that form part of the curriculum. The hierarchy can be used to code responses from a variety of tasks typically included in the curriculum, such as assignments, minute papers, questionnaires and class tests. Although the hierarchy was designed to measure the *development* of students' consideration of variation, it also provides a useful basis for informing an assessment rubric. Furthermore, this qualitative analysis of student responses could be used to develop items that will provide data allowing for a more extensive quantitative analysis of the development of students' consideration of variation.

Future research also should seek to validate the *Consideration of Variation Hierarchy*, by applying it to responses from a wider cohort of students. In addition, the *Consideration of Variation Hierarchy* could be developed and refined further by analyzing responses to a wider range of tasks that include formal inference, more advanced concepts such as experimental design issues, and more complex models such as linear mixed models. For example, a more generalizable hierarchy may include the concepts of systematic and random variation rather than the concepts of within-group and between-group variation, thereby encompassing an even broader perspective of variation.

Longitudinal studies that follow cohorts of students through a number of statistics courses would also inform the further development of the hierarchy. This would then broaden the applicability of the hierarchy beyond introductory courses, providing a more complete picture of the development of students' consideration of variation.

ACKNOWLEDGEMENTS

We would like to thank the editors and the reviewers for their thoughtful comments and suggestions in the refinement of this paper.

REFERENCES

Bakker, A. (2003). Reasoning about shape as a pattern in variability. In C. Lee (Ed.), *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3)*. [CDROM]. Mount Pleasant, MI: Central Michigan University.

Ben-Zvi, D., & Garfield, J. (2004a). Research on reasoning about variability: A forward. *Statistics Education Research Journal*, *3*(2), 4-6.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_forward.pdf]

Ben-Zvi, D., & Garfield, J. B. (Eds.) (2004b). Research on reasoning about variability [Special issue]. *Statistics Education Research Journal, 3*(2). [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2).pdf]

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). [Online: www.amstat.org/publications/jse/v10n3/chance.html]

delMas, R. C., & Liu, Y. (2003). In C. Lee (Ed.), *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3).* [CDROM] Mount Pleasant, MI: Central Michigan University.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). [Online: www.amstat.org/publications/jse/v10n3/garfield.html]

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 397- 409). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J. B., & Ben-Zvi, D. (Eds.) (2005). Reasoning about variation [Special section]. *Statistics Education Research Journal, 4*(1). [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1).pdf]

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 117-148). Mahwah, NJ: Earlbaum.

Garfield, J., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review, 67*(1), 1-12.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education, 10*(2). [Online: www.amstat.org/publications/jse/v10n2/garfield.html]

Lann, A., & Falk, R. (2003). What are the Clues for Intuitive Assessment of Variability? In C. Lee (Ed.), *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3).* [CDROM] Mount Pleasant, MI: Central Michigan University.

Lee, C. (Ed.) (2004). *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3).* [CDROM] Mount Pleasant, MI: Central Michigan University.

MacGillivray, H. (2004). Coherent and purposeful development in statistics across the education spectrum. In G. Burrill, & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable.* Voorburg, The Netherlands: International Statistical Institute.

Makar, K., & Confrey, J. (2005). "Variation talk": Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27-54. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_Makar_Confrey.pdf]

Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course, *Statistics Education Research Journal, 1*(2), 22-37. [Online: www.stat.auckland.ac.nz/serj/SERJ1(2).pdf]

Moore, D. S., & McCabe, G. P. (1999). *Introduction to the Practice of Statistics,* (3rd ed.), New York: Freeman.

Pfannkuch, M. (1997). Statistical thinking: One statistician's perspective. In F. Biddulph & K. Carr (Eds.), *People in Mathematics Education. Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia Incorporated,* (Vol. 1, pp. 192-199). Hamilton: Mathematics Education Research Group of Australasia Incorporated.

Reading, C., & Reid, J. (2005). Consideration of variation: A model for curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable* (pp. 36-53). Voorburg, The Netherlands: International Statistical Institute.

Reading, C., & Reid, J. (2006a). Listen to the students: Understanding and supporting students' reasoning about variation. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, [CDROM]. Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A1_READ.pdf]

Reading, C., & Reid, J. (2006b). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Reading_Reid.pdf]

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Reid, J., & Reading, C. (2004). Just a minute? The use of minute papers to investigate statistical thinking in research, teaching and learning. *Adults Learning Maths Newsletter* (21), 1-4.
[Online: http://www.alm-online.org/Newsletters/News21.pdf]

Reid, J., & Reading, C. (2005). Developing consideration of variation: Case studies from a tertiary introductory service statistics course. *Proceedings of the 55th Session of the International Statistical Institute*, Sydney, Australia. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/13/Reid-Reading.pdf]

Reid, J., & Reading, C. (2006). A hierarchy of tertiary students' consideration of variation, in A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/C122.pdf]

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3).
[Online: www.amstat.org/publications/jse/v10n3/rumsey2.html]

Tippett, L. H. C. (1952). *The Methods of Statistics* (4th ed.). New York: John Wiley and Sons.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal, 12*(2), 147-169.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-265.
[Online: www.stat.auckland.ac.nz/~iase/publications/isr/99.Wild.Pfannkuch.pdf]

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

JACKIE REID
School of Science and Technology
University of New England
Armidale, NSW, AUSTRALIA, 2351

APPENDIX A: CLASS TEST QUESTIONS

LAMPSHELL DATA

Lampshells, although rare worldwide, are quite abundant in parts of New Zealand. Biologists collected a sample of lampshells to see what differences existed between live and dead lampshells. They measured the lengths (mm) of the lampshells. Use the following results to answer the questions given below. (Wild & Seber, 2000)

	Five Number Summary					Mean	Std. deviation
Live	4.12	8.19	15.59	20.37	25.18	14.74	6.61
Dead	10.83	18.27	20.17	22.71	25.93	20.14	3.71

Stem-and-leaf plot: live (N = 40)
The decimal point is at the |

4 | 112889
6 | 2428
8 | 61
10 | 3
12 | 0111
14 | 0256
16 | 896
18 | 170
20 | 00171359
22 | 138
24 | 52

Stem-and-leaf plot: dead (N = 30)
The decimal point is at the |

4 |
6 |
8 |
10 | 8
12 | 48
14 |
16 | 2687
18 | 341999
20 | 1125446
22 | 56734
24 | 36799

(a) Describe the shape of the distributions of lengths for both live and dead lampshells.
(b) Give the appropriate numerical summary for each distribution. Justify your choice.
(c) Compare the two distributions.
(d) Using the IQR, identify any potential outliers for the distribution of lengths for the dead lampshells. Show your calculations.

CAESAREAN DATA

A study in Switzerland examined the number of caesarean sections (surgical deliveries of babies) performed in a year by doctors. The doctors were identified by gender. Use the following results to answer the questions given below.

	Five Number Summary	Mean	Std. deviation
Males	20.0 27.5 34.0 47.0 86.0	41.33333	20.60744
Females	5.0 10.0 18.5 29.0 33.0	19.1	10.12642

Stem-and-leaf Plot: Males (N = 15)
The decimal point is 1 digit to the right of the |

0 |
1 |
2 | 05578
3 | 13467
4 | 4
5 | 09
6 |
7 |
8 | 56

Stem-and-leaf Plot: Females (N = 10)
The decimal point is 1 digit to the right of the |

0 | 57
1 | 0489
2 | 59
3 | 13

(a) Describe the shape of the distributions of lengths for both live and dead lampshells.
(b) Give the appropriate numerical summary for each distribution. Justify your choice.
(c) Compare the two distributions.
(d) Using the IQR, identify any potential outliers for the distribution of lengths for the dead lampshells. Show your calculations.

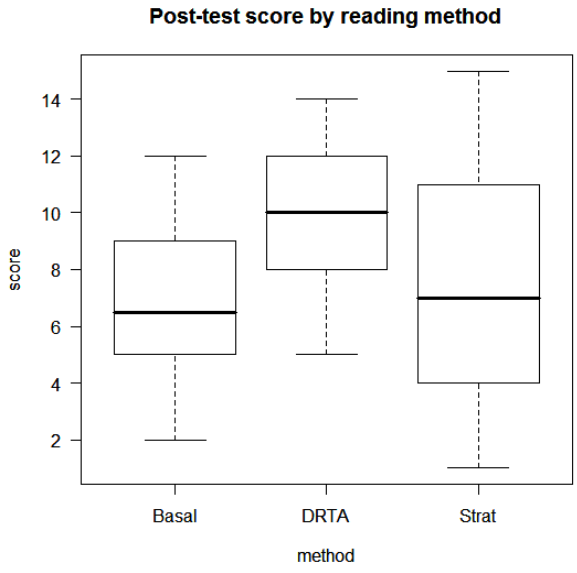
APPENDIX B: ASSIGNMENT QUESTIONS

READING PROGRAMS

Researchers at Purdue University conducted an experiment to compare three methods for teaching reading. Students were randomly assigned to one of the three teaching methods, and their reading comprehension was tested before and after they received the instruction. We would expect no significant difference in test scores between the groups before the teaching methods were used (and that was the case). A measure of reading comprehension for all subjects, from the post teaching period, is included in the dataset.

Reference: Moore, David S., and George P. McCabe (1999). *Introduction to the Practice of Statistics (3rd edition).*
Original source: study conducted by Jim Baumann and Leah Jones of the Purdue University Education Department.
Number of cases: 66
Variable Names:
C1. Group: Type of instruction that student received (Basal, DRTA, or Strat)
C2. POST1: Reading score after receiving instruction using one of the methods.

(a) Summarise the data in a table giving sample sizes, means, and standard deviations. Give an appropriate graphical summary that allows a comparison of the groups.
(b) State and check the assumptions of the ANOVA model:
i) by constructing normal probability plots for each group.
ii) using Bartlett's test.
(c) Give appropriate null and alternative hypotheses to compare the different groups (in words and using statistical notation).
(d) Run the ANOVA, producing
i) a normal probability plot of the residuals.
ii) Tukey's pairwise comparisons.
(e) With reference to the output from (a) and (d), write a non-technical summary of your conclusions.



CUCKOO EGGS

L.H.C. Tippett (1902-1985) was one of the pioneers in the field of statistical quality control. These data on the lengths (mm) of cuckoo eggs found in the nests of other birds (drawn from the work of O.M. Latter in 1902) are used by Tippett in his fundamental text. Cuckoos are known to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds.

That cuckoo eggs were peculiar to the locality where found was already known in 1892. A study by E.B. Chance in 1940 called The Truth About the Cuckoo demonstrated that cuckoos return year after year to the same territory and lay their eggs in the nests of a particular host species. Further, cuckoos appear to mate only within their territory. Therefore, geographical sub-species are developed, each with a dominant foster-parent species, and natural selection has ensured the survival of cuckoos most fitted to lay eggs that would be adopted by a particular foster-parent species.

Reference: L.H.C. Tippett, *The Methods of Statistics* (*4th Edition*), John Wiley and Sons, Inc., 1952, p. 176.

Number of cases: 120

Variable Names:

C1. Length (egg length(mm))

C2. Species (MDW PIPIT: (Meadow Pipit); TREE PIPIT; HDGE SPRW (Hedge Sparrow); ROBIN; PIED WTAIL (Pied Wagtail); WREN)

Is there a significant difference in mean lengths for eggs laid in nests of different bird species?

(a) Summarise the data in a table giving sample sizes, means, and standard deviations. Give an appropriate graphical summary that allows a comparison of the groups.

(b) State and check the assumptions of the ANOVA model:

i) by constructing normal probability plots for each group.

ii) Using Bartlett's test.

(c) Give appropriate null and alternative hypotheses to compare the different groups (in words and using statistical notation).

(d) Run the ANOVA, producing

i) a normal probability plot of the residuals

ii) Tukey's pairwise comparisons

(e) With reference to the output from (a) and (d), write a non-technical summary of your conclusions.