Available online at www.sciencedirect.com
ScienceDirect
Contemporary Educational Psychology 32 (2007) 630–666
Contemporary Educational Psychology
www.elsevier.com/locate/cedpsych

Statistical reasoning of middle school children engaged in survey inquiry ☆

Nancy C. Lavigne [a,*], Susanne P. Lajoie [b,1]

[a] *School of Education, University of Delaware, 206 C Willard Hall, Newark, DE 19716, USA*
[b] *Department of Educational and Counselling Psychology, McGill University, 3700 McTavish Street, Montréal, Que., Canada H3A 1Y2*

Available online 2 November 2006

Abstract

The case study examined two groups of grade 7 students as they engaged in four inquiry phases: posing a question and collecting, analyzing, and representing data. Previous studies reported analyses of statistical reasoning on a single inquiry phase. Our goal was to identify the modes of statistical reasoning displayed during group discussions in all phases as children designed and conducted their own inquiry. A content analysis of audio and video recorded discussions yielded 10 statistical reasoning modes: six relate to Garfield and Gal's [Garfield, J., Gal, I. (1999). Teaching and assessing statistical reasoning. In L. V. Stiff, & F. R. Curcio (Eds.), *Developing mathematical reasoning in grades K-12. 1999 Yearbook* (pp. 207–219). Reston, VA: National Council of Teachers of Mathematics] statistical reasoning types involved in the collection, analysis, and representation of data and four modes deal with an aspect of inquiry not exclusively focused upon in the literature on statistical reasoning—i.e., the problem-posing phase. Although students' reasoning reflected an incomplete understanding of statistics they serve as building blocks for instruction.

☆ This research was supported by fellowships from les Fonds pour la Formation de Chercheurs et l'Aide à la Recherche Doctoral Fellowship (FCAR), Social Sciences and Humanities Research Council of Canada Doctoral Fellowship (SSHRC), and a combined McGill University and Social Sciences and Humanities Research of Canada grant. This work was also funded by a SSHRC grant. Support for the development of the *Library of Exemplars* was provided in part by the Office of Educational Research and Improvement (OERI) through the National Center for Research in Mathematical Sciences Education. The research reported in this paper does not reflect the views of any of these granting agencies.
* Corresponding author. Fax: +1 302 398 4110.
E-mail addresses: nlavigne@udel.edu (N.C. Lavigne), Susanne.lajoie@mcgill.ca (S.P. Lajoie).
[1] Fax: +1 514 398 6968.

0361-476X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cedpsych.2006.09.001

Keywords: Statistical reasoning; Inquiry; Mathematics education; Middle school; Thinking; Cognition

1. Introduction

Research in the area of reasoning is of great interest to cognitive psychologists (Holyoak & Morrison, 2005) and mathematics educators (National Council of Teachers of Mathematics [NCTM], 1989, 2000) alike. This attention is due in part to the role that reasoning plays in problem solving and decision-making in general, and to its key function in the "knowing and doing of mathematics" specifically (NCTM, 1989, p. 81). Reasoning is commonly defined as a "process of drawing conclusions" (Leighton, 2004, p. 3), which is based on how one applies one's knowledge to reach goals in various situations (Evans, 1993). According to Holyoak and Morrison (2005), it is at times difficult to tease apart reasoning from problem solving and decision-making. They explain the overlap this way: "To solve a problem, one is likely to reason about the consequences of possible actions and make decisions to select among alternative actions... Making a decision is often a problem that requires reasoning." (Holyoak & Morrison, 2005, p. 2).

Reasoning in the service of problem solving and decision-making is evident in inquiry situations where the goal is to arrive at decisions that will enable a problem to be solved and where a solution must be produced rather than retrieved from memory (Zimmerman, 2000). In other words, the solution is based on *inferences* that people make from the knowledge they have rather than on their recall of the solution. Two general kinds of reasoning, deduction and induction, play a role in inquiry. Deduction is truth preserving (i.e., inference is made to confirm a hypothesis) and involves reasoning from premises that contain general statements, rules, or scientific laws to arrive at specific conclusions that follow logically from the premises (Holyoak & Morrison, 2005; Leighton, 2004). In the context of inquiry, deductive reasoning is involved in the testing of hypotheses, laws, or theories—e.g., "If my hypothesis is true then I should observe some pattern of evidence" that follows from the hypothesis (Zimmerman, 2000, p. 102). Induction is truth expanding (i.e., the inference leads to new knowledge) and involves reasoning from particular data or observations to arrive at a general conclusion. In inquiry, inductive reasoning is involved in making inferences that produce hypotheses, laws, or theories—e.g., If the data show a particular pattern of evidence then I can make hypothesis X (Zimmerman, 2000).

The mathematics community places a high value on reasoning as illustrated in its creation of a "reasoning" standard (NCTM, 1989, 2000). According to Russell (1999), reasoning is the means by which students learn to understand the abstract ideas that make mathematics the discipline that it is. In this sense, mathematics is about generalizations (Russell, 1999), and generalizations are involved in inductive and deductive reasoning. Both reasoning types are included in the reasoning standard for grades 5–8 (NCTM, 1989)—i.e., students must learn about and use deductive and inductive reasoning and make and evaluate conjectures and arguments. These goals are reiterated somewhat differently in the most recent standards (NCTM, 2000; e.g., select and use various types of reasoning), but the thrust is the same. One recommendation to foster such reasoning is to use problem situations, such as group projects involving the use of technology on problems that are of interest to students, and to augment the complexity by including statistics (NCTM, 1989). In essence, these proposals call for middle school students to reason

collaboratively about statistics in inquiry situations, and for them to use various ways of reasoning to do so.

Deduction and induction are often used in the service of content-based reasoning, such as scientific reasoning (Zimmerman, 2000) and statistical reasoning in inquiry contexts where hypotheses are tested and produced from data. Statistical reasoning refers to "the way people reason with statistical ideas and make sense of statistical information." (Garfield & Chance, 2000, p. 101). Such reasoning includes the ability to make and implement decisions (e.g., about sampling) based on statistical principles (e.g., the law of large numbers [LLN]) and the ability to make inferences from sample data, averages, and graphs to interpret results and draw conclusions (Garfield & Chance, 2000).

Six content specific reasoning types have been identified: reasoning about samples, reasoning about data, reasoning about representations of data, reasoning about statistical measures, reasoning about uncertainty, and reasoning about association (Garfield & Gal, 1999). These statistical reasoning types coincide with activities proposed in the data analysis standard where students are asked to pose questions that involve the collection, organization, analysis, and representation of data to make inferences and predictions in order to answer the questions (NCTM, 2000).

Statistical reasoning can accompany deductive and inductive reasoning in inquiry situations where hypotheses are formulated and tested for experiments or surveys designed to answer specific questions. For example, one might have a hypothesis that being athletic changes one's resting heart rate. We can conduct an experiment and deduce from our theory that if we have matched groups and each group (athletic and couch potatoes) engages in a 15-min activity then the athletic group should have an average heart rate lower than the couch potatoes assuming that they are matched for age, diet, gender, etc. In addition, from this fact we can infer inductively that generally there is something about exercise that leads to changes in resting heart rate. To come to this conclusion, we must decide which statistical test is appropriate to the question, and this decision requires some reasoning about the nature of the tests themselves and the type of data involved. In other words, reasoning about data and reasoning about statistical measures are necessary. Thus, while deductive and inductive reasoning are involved, statistical reasoning is also required for making sense of the data upon which the general reasoning is based, and this sense-making is based on inferences made given one's knowledge of statistics.

Early research on statistical reasoning focused primarily on the first part of the definition (i.e., reasoning with statistical ideas or rules) by examining how abstract statistical rules develop in children (e.g., Piaget & Inhelder, 1975) and the extent to which adults use such rules to make decisions on well-defined hypothetical problems (e.g., Kahneman & Tversky, 1982). Recent research extends the focus to the second part of the definition (i.e., drawing inferences from data) by examining the development of children's ability to reason with data, as well as the extent to which children apply statistical rules in a variety of learning situations. This latter research was motivated by the creation of the NCTM standards (1989, 2000), which espouses data analysis as well as reasoning and problem solving in the K-12 curriculum.

Research findings suggest that children can reason about uncertainty with instructional support (Fischbein, 1975; Jacobs, 1999; Metz, 1998). Moreover, children's ability to reason about data depends on whether they have made the shift from thinking of individual data points (e.g., 10 data points represent 10 individual responses) to thinking in terms of a collection of data (aggregates) that are distributed in a particular way (e.g., averages)

(Cobb, 1999; Hancock, Kaput, & Goldsmith, 1992; Konold, Robinson, Khalil, Pollatsek, & Well, 2002; Lehrer & Schauble, 2004). This capability provides the basis for students to be able to reason inductively, that is, to draw conclusions by generalizing a finding from a set of observed data. It may also facilitate deductive reasoning in cases where students' data collection is driven by a hypothesis that is being tested by the data.

Personal knowledge and beliefs seem to play a critical role in children's statistical reasoning (Jacobs, 1999; Schwartz & Goldman, 1996; Schwartz, Goldman, Vye, Barron, & CTGV, 1998). Similar findings are reported in studies involving inquiry situations (Zimmerman, 2000, 2005). General knowledge and beliefs acquired in social situations can influence the interpretation of formal properties of a statistical problem. For instance, Hancock et al. (1992) found that 15-year old students relied on their personal knowledge about music to reject country and western as an option in a survey on types of music. These adolescents stated that, "nobody listens to that stuff" (p. 357) and their belief was that "nobody should" (p. 357). This example reflects inductive reasoning in that the students are basing a conclusion on a specific case, that is, they are generalizing on the basis of their observations of music preferences. However, the extent to which the content of this reasoning is valid depends in part on whether or not it is based on statistical ideas. The extent to which personal knowledge underlies reasoning in survey contexts and takes precedence over more normative types of statistical reasoning (i.e., reasoning in ways that are consistent with statistical ideas accepted in the domain) is an important question.

Studies with adults (Nisbett, 1993) and middle school children (Jacobs, 1999; Schwartz & Goldman, 1996; Schwartz et al., 1998) suggest that normative statistical reasoning about sampling is more likely elicited on formal statistical or chance problems involving random generating devices than on problems that are more social in nature, such as surveys, where individuals are selected to be in a sample. For example, when the hypothetical sampling context involved M&Ms (i.e., a chance context), Jacobs (1993) found that middle school students who had not received any formal statistical training were able to reason on statistical grounds. However, when students were asked to consider how they would sample individuals (i.e., an everyday survey context), they allowed their prior beliefs of fairness to influence their decision about sampling method. Specifically, students refused to believe that selecting a random sample of individuals from a hat (e.g., Republicans and Democrats) was appropriate because they were concerned that a particular group of individuals (e.g., Democrats) would be left out, which would not be fair and would result in hurt feelings.

Note that personal experience can also influence reasoning in certain chance situations. For instance, children do not always consider dice as fair when they use games as a reference point for thinking about chance. Children believe that numbers that are given special status in games (e.g., 6) are least likely to occur (Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993; Watson & Moritz, 2003). These examples suggest that non-normative statistical reasoning can result from prior beliefs originating in social situations. The extent to which the social situation or prior beliefs play the biggest role in these studies is unclear.

The research findings on statistical reasoning are based on studies reporting on a specific phase of the empirical investigation process, that is, sampling, analysis of data, or representation of data. For example, Bright and Friel (1998) were interested in 6th–8th graders' ability to interpret data presented in a variety of graphical displays (e.g., bar graph, line plot, stem-and-leaf, histogram) and to draw connections between the representations. They presented students with the graphs and examined students' understanding

and reasoning through interviews. A few researchers, such as Lehrer and Schauble (2000, 2004) and Cobb (1999), conducted studies that blend data analysis with data representation. In these design and teaching experiments, students' reasoning about variation is fostered using distribution as the representation. Trying to make sense of the data and drawing conclusions from them inevitably leads to concepts related to data analysis, such as averages. These studies represent the extent of overlap between inquiry phases that we see reported in published work.

To date, few studies have reported on children's statistical reasoning as they make and implement decisions in all phases of investigation (i.e., problem posing, data collection, data analysis, data representation), particularly in the context of surveys. Such research is warranted given the strong emphasis on inquiry in curriculum standards (NCTM, 1989, 2000). Children who engage in inquiry must think about many statistical ideas, and always in relation to a question. This situation enables us to explore the reasoning associated with generating questions, which has been relatively ignored (English, 1998) despite the window that it can provide into students' thinking (Silver, Mamona-Downs, Leung, & Kenney, 1996). Moreover, studies that examine reasoning throughout the inquiry process can provide the impetus for identifying developmental signposts in children's ability to reason *across* inquiry phases. We lack models that describe reasoning at this level. Presently, researchers are studying the development of children's statistical thinking *within* inquiry phases (Jones, Thornton, Langrall, & Mooney, 2000; Watson & Moritz, 2000a, 2000b). For instance, Watson and Moritz (2000a) have explored how children in grades 3–9 developed their understanding of average over a period of 3–4 years (i.e., data analysis phase).

In this paper, we provide an in-depth case analysis of 7th grade students to characterize the statistical reasoning they displayed as they engaged in all phases of inquiry. Students worked in small groups to pose their own inquiry questions and implement the design of their investigation by collecting data, analyzing and interpreting the data, representing the data graphically, and drawing conclusions from the data. We will illustrate the modes[2] of statistical reasoning that students displayed in each inquiry phase—i.e., problem posing, data collection, data analysis, and data representation—when they planned, implemented, and presented their findings to classmates. Later, we discuss how these modes are part of the statistical reasoning types Garfield and Gal (1999) identified as key outcomes for learning statistics.

2. Method

2.1. Participants

We selected six 7th grade students for the case study. The students were comparable to their classmates in terms of demographics. Five students were Caucasian and one student was of East Indian descent. All came from middle-class families and were 12-years of age. Three students were male and three were female. The teacher assigned the students to two mixed groups in terms of gender (i.e., Group A = one male and two females; Group

[2] We use this term to indicate the specific and contextualized ways in which students can reason about statistics. We view modes of reasoning as elements within particular types of statistical reasoning.

B = two males and one female) and elementary schools attended prior to entering 7th grade to provide diversity in groups. Mixed-ability grouping was not possible due to inconsistent grade reporting across elementary schools (K–6) attended prior to beginning the first year at a co-educational preparatory high school (7–11), which was located in a large city. We did not provide students with specific training on how to interact collaboratively, but we strongly encouraged students to discuss their ideas with each other by addressing prompts on a planning sheet.

We selected Groups A and B for an in-depth analysis because the verbal data from the two groups were the most complete compared to other groups in the class. Our goal was to document and describe instances of reasoning across two relatively similar groups rather than to compare contrasting cases.

2.2. Context

Over the course of 2 weeks, we provided students with a framework for thinking about inquiry through modeling and gave them an opportunity to implement this thinking in an investigation that they designed and conducted with classmates. Modeling consists of a more skilled individual carrying out a task so that learners can observe and build a representation of how to solve that task (Collins, Brown, & Newman, 1989). Such observation can facilitate learning (Bandura, 1977). We modeled two types of inquiry, an experiment and a survey, to expose students to a variety of ways in which inquiry can be conducted and to give them a choice of engaging in the type of inquiry that interested them later in the study. Prior to engaging in these modeling activities, the teacher introduced students to the notion of statistics through newspaper articles. In Appendix A we outline the statistical thinking that the teacher and students demonstrated in the introduction, the first modeling activity involving an experiment on pulse rates, and the second modeling activity involving surveys. The manner in which each activity was structured is described next, followed by a description of the students' task after the modeling activities.

2.2.1. Introductory models of statistics presented by the teacher

The purpose of the introduction was to ease students into the inquiry process and to get a sense of their existing knowledge of the term statistics, the inquiry process, and forms of data representations. In the first part of the introduction the teacher asked students what they thought statistics meant. Their responses were that statistics involved "a collection of facts about certain things or numbers," provided "an answer to a question using numbers," and included "scores as well as surveys." The second response reflected a broader conception of statistics than the other two, which dealt with data and data analysis. The teacher prompted students to think about what kind of questions could be asked using statistics, and to consider why data are collected. One student responded to the teacher's query by posing the question "How many people in this room smoke and how many do not?" The teacher broadened the scope of the question to "How many people in Canada are smokers and how many are not?" and asked them to consider who would be interested in the results of the study (i.e., cigarette companies and the federal government given that it covers the costs of health care) and who was being targeted in advertisements (i.e., younger people). The teacher's elaboration emphasized why questions are posed to put the inquiry process in a context.

The second part of the introduction gave the class an opportunity to discuss two questions posed in their local newspaper (a poll regarding a hotly debated political issue and average attendance at football stadiums to support the construction of a new baseball stadium) and the graphical representations used to communicate the results (i.e., tables and bar graphs). The teacher focused students' attention on the questions that were posed by the authors of the articles as well as the authors' choice of data presentation to support their position on the issues. Sampling was discussed using the political question posed in one of the articles. The teacher spent quite a bit of time asking students to interpret the graphs (one presented percentages and the other displayed averages) and to consider whether tables or graphs were more informative in conveying the answer to the questions. The teacher concluded the introduction by outlining the four inquiry phases in the context of the pulse rate activity, which students would engage in on the following day, and by pointing to some of the examples from the previous discussion.

The aim of the introduction was to provide an overview of the inquiry process in the context of questions that could be asked and answered with statistics, rather than an in-depth introduction to statistical content. Students learned the content in the next activity—the teacher-guided experiment on pulse rates (adapted from the American Statistical Association Guidelines for K-12, 1991). This activity was intended to show how descriptive statistics (e.g., variable, sample, population, representativeness, data, measures of central tendency, range, bar graph, pie chart) could be used to answer a question posed for inquiry. The pulse rate investigation posed the following question: Do the at-rest pulse rates of grade 7 students differ from their pulse rates after physical activity (i.e., jogging in place for 5 min)? The class decided that its population would be grade 7 students in the province of Québec and that the sample would comprise the entire class. The teacher indicated that this was not a particularly good sample because it was not sufficiently diverse. Nonetheless, this decision was made in light of time constraints. The teacher introduced the concepts of variable, sample, population, and representativeness on the basis of the question. Students were asked to make a prediction and then to conduct the experiment.

After the data were collected, the teacher showed students how to organize and enter the data into a spreadsheet in a way that would enable them to compare the before and after exercise pulse rates. Measures of central tendency (mean, median, and mode) were introduced as students considered how to summarize the data for interpretation. A discussion about which measure provided the most accurate finding led to the conclusion that the mean and median were better than the mode for the pulse rate data that the class had collected (the data were bimodal), even though there was an outlier in the data that affected the mean. Graphs were explored as a way to represent the data, and conclusions were drawn on the basis of analysis and representation. An issue that the teacher raised was whether each individual data point or average should be represented. Graphs that presented both were constructed.

2.2.2. Models of statistical investigation facilitated with technology: library of exemplars

To illustrate statistical thinking in a different inquiry context, students were shown examples of surveys that previous students of the same age had presented to peers. In our research we have used technology as a medium for providing exemplars of student work through digitized video[3] (Lajoie, 1997, 1999; Lajoie, Lavigne, Munsie, & Wilkie,

[3] We did not have examples of experiments at the time the LE was developed.

1998; Lavigne & Lajoie, 1996). Since students would be evaluated on their inquiry, textual descriptions of evaluation criteria supplemented the video clips and were provided on the computer in what we refer to as the library of exemplars (LE, Fredriksen & Collins, 1989). The intent was to give students good (referred to as appropriate) and better (referred to as more appropriate) examples for each phase of statistical investigation to stimulate a discussion of the criteria.

A brief description and two examples were presented in the library for each inquiry phase—i.e., posing a question, data collection, analysis, and representation. For instance, the posing a question criterion consisted of presenting a clear question, specifying the variable, categories (where applicable), and population of interest. One example illustrated a general or open-ended survey question, ''What is your favorite fastfood restaurant?'' The second exemplar showed the same student who restated the group's question more precisely, ''What is your favorite fastfood restaurant between McDonald's, Harvey's, Wendy's, Lafleur's, Burger King, and other?'' The teacher asked the students which question was better according to the criteria. The students decided that the forced-choice question (the one with categories) was better than the free-response question (without categories) because it specified what the group wanted to find out (i.e., it met the clarity criterion in terms of specifying survey categories) and avoided the problem of obtaining an indefinite number of free responses. The inclusion of the category ''other'' was desired because it eliminated dishonest responses should someone's preference not be reflected in the categories that were chosen. The teacher supported the students' decision that the forced-choice survey was clearer. At the same time the discussion went beyond the need for developing a clear research question and emphasized the importance of the clear choice categories versus open-ended questions.

The LE exemplars focus on a specific element in a criterion (or inquiry phase) rather than addressing a full range of issues that apply to each criterion. The intention was to promote discussion and a way of thinking about certain aspects of inquiry within a phase. The students' task in the LE was to read each description, view the video clips, and discuss whether or not the examples met a criterion. They did so in groups for each criterion and then shared their ideas with the class. Students viewed the question posing and data collection criteria and exemplars on one day, discussed which exemplars were better, and then discussed possible questions that their group could explore and the manner in which they would collect that data. On the following day students observed the data analysis and representation examples, discussed which exemplars were better and how they could be improved upon, and then addressed these issues as they planned their group investigation. The teacher's task was to extend students' thinking by modeling questions about the quality of the exemplars presented in the library.

2.2.3. Student-generated statistical investigations

Students embarked on their own statistical inquiry after completing the modeling activities. The students' task was to develop any type of inquiry that their group was interested in conducting—an experiment or a survey—and to engage in four phases of statistical inquiry—problem posing or developing a research question, data collection, data analysis, and data representation—using knowledge of descriptive statistics. Each group was given two sets of planning prompts that they had to respond to verbally and in writing. The first set of prompts reminded students of the issues they needed to address in generating a question and in collecting the data. The second set of prompts focused on issues pertaining to

data analysis and representation. All prompts restated the criteria that were presented in the LE as questions and are displayed in Table 1. All activities took place in the school's computer lab during students' regular mathematics class. Students worked in small groups at a computer workstation using ClarisWorks™ software to enter, analyze, and graphically represent data (Claris Corporation, 1992). The teacher monitored each group by visiting each of them and addressing their questions about the criteria.

Two issues need clarification at this point: (a) the similarities and differences between the two types of inquiries modeled (pulse rate experiment and surveys in LE) and (b) how closely aligned the modeling activities reflected what was expected of students on the group inquiry task. To address these issues we need to contrast the concepts and inquiry phases emphasized in the two activities and the concepts and phases that students were expected to address in their group investigation. The contrast is presented in Table 2. Since the planning prompts structured the students' task during group inquiry they provided the means for examining alignment to the modeling activities. Table 2 outlines the concepts that the planning prompts focused upon in each phase (e.g., RQ refers to Research Question) and how these concepts were addressed the modeling activities.

Table 1
Inquiry planning prompts

Inquiry phase	Prompts
Research question	Decide on your group project. Write down the research question that you will investigate in (a). Remember to make your question clear. Then answer (b), (c), and (d). (a) Research question: (b) What is the variable in the question? (c) Are there any categories associated with it? If so, what are they? (d) What is the population you are trying to investigate?
Data collection	How is your group going to collect the data to answer this question? Consider the following issues and write down your answer in the space below: (a) How large will your sample be? Why did you choose this size? If your sample divides into groups, what is the size of these groups? Why did you choose this size? (If your question does not involve groups, you do not have to answer this question) (b) Where are you going to collect the data? Why? (c) How are you going to make sure that your sample accurately represents the population you are seeking information about? (d) What type of data are you going to collect? (e) How are you going to collect the data?
Data analysis	Decide on how your group will analyze the data you collected to answer your research question. (a) What type of statistics will you use? Why? (b) How do you think this analysis will allow you to interpret the results? (c) Would your findings change if the study was done differently? Why or why not?
Data representation	Think about how you will represent the data. (a) Show how you plan to organize the data in the computer spreadsheet (i.e., before you do a graph). Show how it would look like in the spreadsheet here. (b) How do you plan to represent your findings (i.e., in the form of data, tables, and/or graphs)? Why? If you chose a graph, what type of graph best shows the answer to your question? Why? (c) How do you expect to describe the findings and the representation (s) in the oral presentation? In other words, what do you plan to say to your classmates? (d) How do you plan to explain what your findings mean (i.e., how does it answer your question and what does it say about the population addressed in the question)?

Table 2
Alignment of planning prompts guiding statistical inquiry with modeling activities

Concepts in planning prompts	Pulse rate activity	Library of exemplars
RQ: Variable	Pulse rate	Favorite fastfood restaurant
RQ: Category	Not applicable	McDonald's, Harvey's, Burger King, Wendy's, Lafleur's, and other
RQ: Population	Grade 7 students in Québec	Results would differ for people in Toronto
DC: Sample and group size	Grade 7 students (27). Groups not discussed (there were 2 groups: before, after)	50 children because they eat fastfood. An example of groups was not given and thus not discussed.
DC: Location	The high school	Mall or school
DC: Representativeness	Not representative but convenient	If done in one school not particularly representative
DC: Type of data	Not discussed	Not discussed
DC: Sampling method	Not discussed because sampled an intact group but the teacher did randomly select groups to present their arguments in class.	Variety of children at a mall (no specific selection method mentioned) or take a certain number in each grade if done in a school
DA: Type of statistics	Mode, median, mean	Mean, median, mode, range
DA: Facilitate interpretation?	Mean and median best	Not discussed
DA: How would results change?	Not discussed	Results would change if sample were from another province where local fastfood restaurants might be more popular
DR: Data organization	Two columns (before, after) and data is sorted to find the median	Not discussed
DR: Which representation?	Bar, stacked bar, and line using individual data points and average. Chose the line graph with sorted data points as best.	A table and 4 tiled pie graphs comparing the data collected by each group member were presented. The pie graphs were thought to be best.
DR: Describe findings and representation	The line graph of sorted data hides the outlier	Color as an interesting feature of pie graphs
DR: Explain findings	The after pulse rates were higher than the before—by 25	The findings displayed in the table were not discussed at all

The inquiry phases and the content associated with these phases did not substantially differ in the two types of inquiry modeled to students. The main differences between the pulse rate experiment and the LE surveys were the type of questions posed (and thus the purpose of inquiry), the type of data collected, and the measures appropriate for analyzing each type of data. Conducting a survey meant that students had to consider whether or not to include options (or categories), a decision that was unnecessary in the pulse rate experiment. The teacher allotted quite a bit of class time to the discussion of the survey question exemplars in the LE precisely for this reason. The second key difference between the two modeling activities, type of data, was never addressed in the modeling activities. However, this omission did not hinder students in their ability to gather data or to draw inferences from either numerical data (i.e., pulse rates) or categorical data (i.e., responses to surveys).

Finally, students did get the sense that certain measures of central tendency were more appropriate to certain kinds of investigations. The teacher explicitly told them that the mean and median were the most accurate measures for the pulse rate data (note that considering the outlier in the data may have led to a different conclusion). Although the teacher did not indicate that the mode was more appropriate for surveys of the kind discussed in class; his definition of this term (i.e., most frequent number) and his acceptance of the use of the mode in the survey examples is an implicit acknowledgement of its viability in this situation.

To ensure that the differences between the pulse rate experiment and the LE surveys did not negatively impact students' thinking when they engaged in their own group inquiry, we examined the transcripts for instances in which students inappropriately transferred knowledge gained in the pulse rate activity to the group survey. There were only two instances of this type of transfer and both had to do with averages. Specifically, one student insisted that the frequencies be rank ordered even though the median was not the measure used because this procedure was employed during the pulse rate activity. The other instance involved another student who wrote that the mean would be most appropriate to the frequency data because it was accurate. Indeed, this was the teacher's conclusion after analyzing the pulse rate data (although the median was considered equally valid). These errors were not problematic since both students went on to analyze the data in terms of the mode. However, such errors need to be avoided in future.

Table 2 shows that all of the content referred to in the planning prompts were addressed in one or both of the modeling activities, except for two concepts: Type of data and group size. From this information, we conclude that the statistical concepts and inquiry processes learned in the modeling activities were fairly well aligned to the concepts and processes students were expected to apply in a group inquiry task.

2.3. Measure

Statistical reasoning was examined during Group A and B's discussions as they designed and implemented their survey and orally presented the results to classmates. Groups were urged to think about statistical issues in their planning by responding to written prompts orally and in writing before carrying out each inquiry (see Table 1). Earlier work suggests that structure is needed to guide meaningful student inquiry (e.g., Petrosino, 1998). Groups formulated their own question in class and they had to specify their variables for their experiments or surveys. For surveys, they had to decide whether or not to generate categories for the variable. Sampling ideas were generated after decisions were made about the survey question. Data were then collected outside of class followed by the creation of analyses and graphical displays. Groups had 4 days to complete their investigation and the teacher provided guidance throughout the process. Each group orally presented its statistical survey to the class using a computer and projector and was then evaluated by peers, the teacher, and the group itself.

3. Fidelity to the modeling guidelines

The classroom teacher introduced the topic of statistics and implemented the modeling framework on the basis of written guidelines that we gave him on how to conduct the activities and which aspects of statistical thinking to model. To determine the extent to

which the instructor followed each guideline, we viewed the videotaped sessions and utilized a checklist to indicate whether or not guidelines were implemented. We found that the teacher applied 70.49% of the guidelines. Omitted guidelines involved sampling, measures of central tendency and minor issues. While the teacher generally addressed what made a good sample and gave an example of an unrepresentative sample (i.e., asking people in a completely Anglophone community whether the province of Québec should separate from Canada) he did not elaborate on ways in which sampling bias is reduced (e.g., through randomization or sample size). In addition, particular problems associated with each measure of central tendency (e.g., outlier influencing the mean, bi-modal distributions) were not discussed. The remaining omissions consisted of not elaborating on examples and not discussing different types of data that can be collected. The description of each modeling activity provided previously is based on what the teacher modeled.

4. Coding of the data

The main source of reasoning data consisted of audio and video taped recordings of group dialogues during the production and presentation of an investigation, as well as written responses to the planning prompts. The data were coded in seven steps. First, the verbal data from audiotapes were transcribed based on Bracewell and Breuleux's (1994) conventions. Second, the accuracy of audio transcriptions was verified using the videotapes of group dialogues. Written responses to prompts on the design sheets were added to the transcript where appropriate to provide the most complete picture of reasoning as possible. Third, transcripts were segmented based on change in speakers. When a student spoke at length, his/her dialogue was segmented into idea units (Chi, 1997).

Fourth, reasoning segments were identified on the basis of explanations that students provided in response to prompts and to each other to justify their point of view in proposing an idea and in rejecting another's idea. The conclusions that students drew from data either through eyeballing the spreadsheet or interpreting the graph they constructed were also counted as reasoning. Explicit and implicit reasons were coded. We used the following types of explicit markers: causal (e.g., so that, because, therefore), conditional (e.g., if-then, and but), and analogies (e.g., like, similar to). A segment, such as "we want to collect more data because we're missing information" contains the marker "because" and thus would get coded as reasoning. In contrast, implicit reasons are not marked by key words. Below is an example using fictitious names.

Jamie: Let's do 40 students.
Cathy: No.
Jamie: It balances out, 20 for each grade.

In this example Jamie provides a reason for his proposal after a group member disagrees. Although he does not use a causal connector, it is implied—i.e., "Yes let's do [collect data from a sample of] 40 students *because* it balances out, 20 for each grade." Both of Jamie's utterances would be coded as reasoning since they form the explanation; however, they would count as one instance of reasoning since together they constitute the explanation.

The fifth step of data coding consisted of performing a content analysis to create reasoning categories that encapsulated the statistical thinking that was demonstrated.

Although there were examples of general reasoning, such as students negotiating who would collect what data based on time constraints (e.g., Jamie will go out and collect data from the elementary school because he has the time whereas I do not), our objective was to document the ways in which students reasoned statistically. Hence, we did not code instances of general reasoning (i.e., content-free reasoning) in this paper.

Statistical reasoning categories were generated by (a) reading transcripts for each group, (b) identifying reasoning segments in each transcript, (c) identifying the inquiry phase that was the focus of a reasoning segment, (d) identifying the main statistical idea conveyed in each segment (e.g., representativeness, variability, data organization, interpretation), and (e) creating a code reflecting the main statistical idea. The reasoning segments were coded on each transcript to preserve contextual information necessary to make sense of group discussions and reasoning. The coding scheme was developed iteratively. Initial codes were created and applied to transcripts for one group, elaborated and refined, and then applied to the larger set of verbal data with further refinements.

The sixth step consisted of validating the coding categories. A colleague who was not involved in our research was enlisted in the validation process. Her qualifications were that she had done previous work on children's statistical thinking, served as a statistics consultant for many years, and had extensive experience in coding verbal transcripts of children's thinking. We provided her with information on what students were asked to do (e.g., description of task, a copy of the planning prompts), the coding manual with information on transcription conventions and the coding categories, and the coded transcripts. Her task was to go through all of the transcripts and to determine whether or not (a) the reasoning categories reflected the statistical ideas conveyed in the coding scheme (i.e., were the reasons accurately characterized?) and (b) the coding scheme captured all of the statistical reasoning that was demonstrated (i.e., were all instances of statistical reasoning documented?). Changes were made to the coding scheme as a result of this process.

The final step of data coding consisted of establishing inter-rater agreement between two raters (the first author and a discourse analyst[4]) who coded all of the transcripts independently. Reliability was determined in two stages. First, the second rater identified segments in the transcripts that consisted of reasoning. Inter-rater agreement was calculated by dividing the total number of segments that the two raters agreed upon by the total number of segments agreed upon plus the total number of segments with disagreement. Overall agreement was 75.15%. Discrepancies were resolved after discussion. Second, the second rater assigned reasoning types to the reasoning segments using a description of the reasoning categories and a list of the planning prompts that the groups used to conduct their inquiry. Overall agreement before discussion was 83.47%. Agreement on specific reasoning types were the following: population relevant = 65%, variety-based = 94%, category-level = 74%, reasoning about alignment to question = 84%, LLN oriented = 61%, characteristics = 92%, frequency-based = 38%, standardization = 90%, organization-based = 91%, interpretation-based = 94%. Disagreements were resolved through discussion.

[4] The second rater was not the same person who validated the reasoning categories.

5. Results

To provide some context for the results on statistical reasoning, we describe the inquiry project that each group produced and characterize students' understanding of key concepts invoked by the inquiry planning prompts. We outline this understanding separately from reasoning because groups often made decisions that reflected comprehension but the reasons for the decisions were not explained. This section is followed by a description of the modes of statistical reasoning displayed during the inquiries. Finally, we conclude with a description of how the modes of statistical reasoning demonstrated can be traced to the introduction and modeling activities, and to the teacher's scaffolding during the inquiry process.

5.1. Group inquiries produced and students' understanding of key concepts

Both groups produced surveys, responded to prompts on the planning sheet, and sought the teacher for clarification on the prompts or for advice on surveys. An overview of each group's decisions is presented in Table 3. Each group's decision and the understanding of concepts underlying these decisions are described next.

5.1.1. Group A

The group quickly decided that the survey would involve pets. However, the specifics of the question changed through students' discussion with each other and the teacher. Ideas were as follows: how many people in this school have a hamster, how many people have a pet, how many people have pets that died in less than 12 years, and how many grade 7 students in this class have more than 2 pets or how many have more than 3 pets? When the teacher heard that the group intended a survey on number of pets (e.g., 2, 3) he suggested that they make the question more general and pushed them to think about the value of a question by asking them the following: Would you not also want to know about

Table 3
Overview of surveys produced by groups

Inquiry phase	Group A	Group B
Research question	How many grade 7 students in this school have pets and how many don't? If so, what kind?	Does profession change (helper, entertainer, scientist, builder, businessman, other) from grades 7, 9, 11?
Data collection	Considered sampling 50 grade 7 students. Sampled 20 by a list to avoid duplication. Asked them during lunch. Planned to ask an equal number of boys and girls.	Considered sampling 10 students per grade. Sampled 21 (10, 8, 3 making per grade) by asking anyone in each grade. Planned to ask the same number of boys and girls.
Data analysis	Percentages to determine the most popular pets (i.e., mode)	Planned on the mean but used percentages
Data representation	Column graph showing number of dogs, cats, fish, and birds with "other" and "none" categories. A pie chart identifying the "other" pets (i.e., gerbil, turtle, hamster) and main pet types.	Difficulties in organizing data (e.g., alphanumeric entries rather than numeric). Created three pie charts; one per grade.

students who only have 1 pet? Why focus on just 2 pets? Why not generalize it to how many pets? After this exchange the group decided to ask the question "how many pets do grade 7 students have?"

Group A's survey was distributed to grade 7 students in their school and although the question focused on the number of pets, students identified the kinds of pets that respondents reported in answering the question. These data enabled the group to answer two questions: the planned question of how many pets and the question of which kinds of pets that emerged after the data was collected. This situation created a lot of confusion for one student when the data had to be organized to answer the question. She did not know which question the group was answering. The group understood what its variable was, identifying it as "number of pets" for the initial question. Since the variable was only discussed during the initial planning phases, the group did not revise it subsequent to changes to the question.

The students' grasp of population was not strong. They knew that the population in their survey consisted of grade 7 students. However, when discussing population their discourse implied that they were thinking of the sample. Initially, a student had suggested that the "population" was 50 students (25 girls and 25 boys) after finding out that this was the total number of grade 7 students in the school. This suggestion suggests a grasp of the concept. However, the group's decision that their population consisted of 15 girls and 15 boys (30 students) indicates that the concept was not understood. The emphasis on a subset of individuals from a larger group is more consistent with a sample than a population. The group did understand the term *categories* as it explained that it involves giving "a choice, like, do they say what number of pets between a cat and a dog sort of thing" and specified that there were none in its survey. Indeed, the group's question was open-ended requiring free responses from respondents.

When prompted about sample size, Group A planned a sample size of "15 boys and 15 girls," which was the same response it gave in addressing the population prompt. This statement confirms that the part/whole relationship of the sample with the population was not grasped. However, there were some interesting elements in their discussion. One student made a subtle distinction between sample size and sample characteristics (i.e., gender). He commented that, "the sample is the amount of people that we're asking though. It's not who we're asking [gender characteristic]."

Students in Group A also displayed some general knowledge of sampling that can provide the building blocks for a deeper understanding. They wanted to be systematic in collecting the data. For instance, students knew that asking a respondent to participate in the survey twice was not desirable. To avoid this problem, the group decided that each member would ask respondents if they had previously participated in the survey, and if so, to not include them in the sample. Another suggestion was to generate a list of potential respondents prior to gathering the data and to assign the same number to each person in the group (i.e., 10 people each). This idea was extended to dividing each person's list by gender so that each group member surveyed the same number of boys and girls (i.e., 5 of each).

Even though students were systematic in planning for data collection from a list of participants, reducing duplication of efforts, they did not grasp the notion of random selection. Although the concept of randomness was not directly taught, the teacher did refer to the term when he used a calculator to randomly select groups to present their arguments during the modeling activities. When the teacher asked the group how it was going to sam-

ple individuals, students said "randomly." Students did not explain what they meant by random and what procedures they would use to ensure random selection; however, the manner in which the group intended to gather the data shows that the students did not understand it. For instance, the group relied on whom they knew or on who would be in the next class when sampling. The group ensured that an equal number of boys and girls were represented in their sample, a procedure that is not random selection. Students never explained this decision but there seemed to be an implicit agreement that this method was justified.

Another concept that was not understood was group size (i.e., number of comparison groups). One student in Group A indicated that the group size was 15 boys and 15 girls while another took it literally and said that the group size was three because there were three members in the group. If gender had been a comparison group then the first student's response would be related to the concept, but it was not.

Finally, the data revealed that students were confused by the word "statistics." Students thought that statistics encompassed the four inquiry phases that were presented in the teacher's introduction to statistics. The teacher clarified by explaining that in the planning prompt statistics referred to the computation of statistics (e.g., averages to analyze data) rather than the overall process of statistical inquiry. Students did seem to grasp the different types of averages, at least procedurally. For instance, one student defined the mean as "adding up all the numbers and then dividing by how many." For its survey, Group A presented the mode (i.e., the most frequently occurring pet) using two graphs: a bar chart showing the frequencies in each pet category, one of which was "other," and a pie chart showing the percentages in each category with the "other" category broken down. The concept of data type was not addressed in the group discussions probably because it was not discussed in the modeling activities. However, "numbers" was the written response to the prompt on the group's planning sheet. This student was likely thinking of frequencies (and thus a number) although technically, the data were categorical.

5.1.2. Group B

A variety of ideas were generated for Group B's survey, such as how many people read books (i.e., children's books, fiction, etc.), what do children do to relax, what percent of people watch television, what type of music do you like (i.e., classical, rock, etc.), what are your favorite classes (i.e., French, geography, drama, math, etc.). Considerable discussion revolved around this last question but it was rejected after the teacher urged the group to generate a question that it truly wanted to investigate. After this intervention, one student suggested the question "what do you want to be when you grow up," which evolved into the question the group surveyed. This question was selected because two of the students felt that it was the best question out of all the ones generated, although they did not explain why.

Group B conducted a survey on profession preferences and while selection categories had been generated, they were not used during data collection. The group understood that the concept of variable describing it as "the profession of choice," and even though categories for this variable were planned, the group did not refer to the options as categories in discussions or on the planning sheet (which could explain why they forgot these categories during data collection). Group B initially considered students from grades 1 through 11 in two schools (one for grades 1–6 and another for grades 7–11). However, due to time constraints and logistical problems, sampling from this population was not possible. This

obstacle was initially quite frustrating to students in the group who wanted to go all out with the survey. It actually worked out well for the other group member who wanted a simplified question because she did not want to go to the elementary school to gather the data. The group changed the population to students in grades 7–11 in their high school after they realized that it was "okay" to revise their question.

Students in Group B did grasp the part/whole nature of the relationship between a sample and its population. Like Group A, Group B wanted to sample an even number of boys and girls even though gender was not a comparison that was planned or made and the group did not explain why doing so was necessary. The number of students who were included in the sample was small but there was a disagreement about whether or not it should be larger. Two students in the group wanted more respondents in the sample explaining that a larger sample results in more accurate results. However, the third group member was driven by practical concerns and wanted the sample size to be manageable. The manner in which respondents were sampled was not random, even though the group indicated that they had randomly selected its sample during its oral presentation. Each group member selected five girls and five boys by visiting advisor groups across grade levels that meet at specific times and surveyed "whoever was there and if they were in the appropriate grades we chose them." The problem is that the group did not explain how the advisor groups were selected. Nonetheless, Group B did have some knowledge of sampling that could be used to build upon in future instruction.

Like Group A, students in Group B did not have a sense of the terms group size, data type, and statistics. Group B did have a comparison group—i.e., grade levels. However, they did not discuss this planning prompt in discussions or provide a written response on the planning sheet. Students did not recognize that group referred to grade level. Data type was described as being "data about what you're going to be when you grow up and how it changes as you get older." This is a restatement of the research question and categories are implied, but we cannot be sure that this is the manner in which it was understood without more evidence. In addition, students in Group B had to ask the teacher to explain what was meant by "statistics" when it came time to discuss data analysis. There was no discussion of the measures of central tendency in this group. However, students did provide a written response to the statistics planning prompt that the mean would be used for the survey because it is accurate. Why the mean was thought to be more accurate was not explained; however, the mean is not the most appropriate measure for frequency data. In the end, the most frequent response (mode) was used for analyzing the data. The group constructed three pie charts, one for each grade level, depicting percentages. Although pie charts are a viable option, a bar graph comparing responses to each category by grade level would have been more informative.

5.1.3. Summary

Even though Group B had planned to conduct a survey with categories, it investigated a survey with an open-ended question (i.e., free response) just as Group A did. Both groups made similar sampling decisions. The sample size tended to be small and sampling occurred in the school either by making a list of potential respondents or by asking anyone who happened to be at the sampling location. Both groups insisted on sampling an equal number of boys and girls even though gender was not a comparison variable. The reasons for this decision are unknown and appeared to be an assumption that was not questioned by students in either group. Finally, both groups used percentages to show the mode

through a pie chart, and Group A included a bar graph representing frequencies. The students did have a basic understanding of key concepts while a couple of concepts that were not addressed in the modeling activities were clearly not grasped (i.e., group size and data type).

5.2. Modes of statistical reasoning demonstrated in survey investigations

Statistical reasoning categories emerged during discussions centering on four issues: (a) whether to construct a forced-choice survey with selection categories that the group would need to establish prior to sampling or whether to pose an open-ended question that would allow respondents to give their own answers; (b) which categories to include in the survey if a forced-choice question is used and how many; (c) how many respondents should be sampled; and (d) how to group responses to an open-ended survey question into categories subsequent to data collection. Ten modes of statistical reasoning were identified through discussion of these issues: population relevant, variety-based, category-level, Law of Large Numbers (LLN) oriented, characteristics oriented, frequency-based, organization-based, interpretation-based, alignment to question, and standardization oriented.

The frequency with which each statistical reasoning mode occurred during discussions in each inquiry phase (identified by prompts) is displayed in Table 4. These data show that most of the reasoning occurred during the analysis and representation inquiry phases (115 reasoning segments) followed by problem posing (64 reasoning segments). The reasoning during data analysis and representation was mainly organization-based and interpretation-based, both of which were evenly distributed across groups. The high incidence of category-level reasoning in this inquiry phase was due to the fact that both groups collected nominal data as a result of investigating an open-ended survey. Population relevant reasoning and variety-based reasoning occurred most frequently during problem posing and were displayed mainly by Group B because they discussed categories for two forced-choice questions (i.e., favorite subject and profession choice) while Group A only briefly considered a forced-choice question. The less frequently reasoned about inquiry phase was data collection (31 reasoning segments) where both groups displayed the same amount of reasoning.

Table 4
Frequency of occurrence of reasoning types per inquiry phase

Modes of reasoning	Problem posing	Data collection	Data analysis/representation	Total
Population relevant	19	5	0	24
Variety-based	22	2	2	26
Category-level	10	2	33	45
LLN oriented	2	11	5	18
Characteristics	4	3	17	24
Frequency-based	0	0	10	10
Standardization	4	5	1	10
Organization based	1	3	60	64
Interpretation based	0	0	41	41
Alignment to question	2	0	28	30
Total	64	31	115	292

Of particular note is that each statistical reasoning mode was most frequent in a specific inquiry phase (e.g., population relevant and variety-based in problem posing, and interpretation-based in data analysis and representation) except for standardization-based and category-level reasoning. Still, the fact is that statistical reasoning modes cut across inquiry phases, which illustrates the recursive nature of inquiry where ideas or decisions are revisited when students consider multiple phases.

Having outlined the frequency with which statistical reasoning modes occurred and their distribution across inquiry phases, we now describe the modes.

5.2.1. Reasoning about population relevant survey categories

Population relevant reasoning refers to proposing and selecting categories for a forced-choice survey based on perceived relevance to the interests of the population being surveyed. Groups indicated that the population of interest in all surveys consisted of students in the school. When discussing categories they considered the interests of students in general, an indication that they were focusing on the population rather than on the specific sample, which consisted of students in a particular grade. This mode of reasoning is often guided by students' personal knowledge as they choose peers as the population (e.g., we are grade 7 students and we love pizza so our peers are likely to as well, so we should include this category in our survey). Group B specifically displayed population relevant reasoning in discussions of a possible survey question. For example, Jamie rejected Alex's suggestion of sports as a category because he did not feel that it represented general professional interests (i.e., "because some people are not into sports") and suggested that astronaut be added because it reflected the interests of young children (i.e., "whenever you talk to little kids, it's like I want to be an astronaut when I grow up. It's true.").

5.2.2. Variety-based reasoning

Variety-based reasoning occurs when dialogue centers on how many questions or categories to include in a survey. The goal is to represent the diverse interests of the population by providing more than one question or a range of categories because there is a concern that all respondents will choose the same option (or that they will all say "yes" to a question) because there are no other options. In other words, students want the survey to elicit different responses. This same reasoning applies to choosing a question that yields different responses from comparison groups. An example of variety-based reasoning is provided by Jamie when he discussed a question, "What's your favorite subject?" with his group (B). Jamie seemed worried about the responses that the question would generate. For example, he asked the group "you don't think everybody is going to say the same thing?" and then suggested two questions, one on academic subjects and one on nonacademic subjects. This idea was rejected. Later, Cathy and Alex decided on math, gym, and drama as categories. Again, Jamie asked, "You don't think we should have two questions? 'Cause I doubt anybody will say I like math more than drama!" In other words, he anticipated all of the responses to be drama.

5.2.3. Category-level reasoning

Category-level reasoning can occur during problem posing or data organization. In both cases the issue involves grouping—i.e., categories or responses. Jamie (Group B) provides an example of category-level reasoning during problem posing. The group had decided to conduct a survey on professions that students in different grades wanted to pursue in

the future. The categories for types of professions were generated (i.e., entertainer, scientist, helping profession) along with specific examples of each (i.e., entertainer = musician, actor; helper = doctor, policeman; scientist = scientist, astronaut). When Cathy suggested removing the examples, Jamie disagreed saying, "'cause they won't know what we're talking about, you can't say scientist, they might not know we're talking about astronaut." In essence, Jamie argued against using only one kind of profession as categories, which is a basic level concept, and made a case for including specific professions, which are subordinate level concepts. Jamie did not want to lose information about what makes an astronaut and a scientist distinct.

Category-level reasoning was also evident in dialogues during data representation as groups discussed how to organize responses from an open-ended survey into categories. For example, Joly (Group A) wanted to represent every type of pet that was identified in the first survey (i.e., how many grade 7 students in this school have pets and how many don't? If so, what kind?) rather than group some categories into an all encompassing category called "other." Her reasoning was that it was more accurate to be able to identify the actual types of pets. For example, at Dana's suggestion that some responses be grouped as "other" Joly replied, "No, 'cause I need to know if the other is really an "other" or not. I want to know what the other really is... 'cause that's being more accurate than other." According to Joly, a pet survey was accurate only if it represented all pets identified, and since the category "other" lacks pet-like features it should not be used. Evidence for this idea is provided when Joly responds to Dana's suggestion that respondents with no pets (i.e., none) should be added to the "other" category:

> Are we doing "none" as "other" animals, which means everybody in the whole entire world has animals and we're not mentioning the "other" animals or are we doing ... or are we doing the other way where we're telling what everybody has and if they have none they have none, and instead of having an animal which doesn't really exist in the world [category "other"]?

5.2.4. Law of large numbers (LLN) oriented reasoning

LLN oriented reasoning involves making sampling decisions based on a primitive version of the statistical principle that the larger the sample size and the less variable the sample, the more accurate the inference is from a sample to a given population. Jamie and Alex's (Group B) statements that "more is better" and that "the more the people, the more accurate" in discussing sample size indicate that they were aiming for a large sample to enhance the accuracy of their results. Although the students did not use the term representativeness, the notion that a larger sample is better or more accurate than a smaller sample is related to the LLN. The consequence of a small sample for drawing conclusions was acknowledged in discussing the results. Jamie said to his peers "but we only did three so this can't be that accurate" and elaborated to classmates in the oral presentation: "if we did this like, this project again when there were actually more than three grade 11s there and they weren't all on the Stratford trip, we'd probably get a different result than 100% for businessman 'cause they were like, not many people there." Drawing conclusions in light of limitations involves coordinating sampling and data and is a critical step in completing the inquiry cycle, one often overlooked by students.

5.2.5. Characteristics oriented reasoning

Students demonstrated characteristics oriented reasoning when they referred to different characteristics of samples in reasoning about sampling and whether or not the results would change if the study were conducted differently. This type of reasoning is primitive in that students do not consider the bias in their sampling procedure, which reduces the representativeness of the sample to the population, but instead think about characteristics of a new sample. Jamie from Group B provides an example of this type of reasoning when he explained to his classmates in the group's oral presentation, "Yeah, but if we did it somewhere else then it would be, things would be different 'cause I mean, say we did it is some little like village, they'd want to be like a farmer or something like that." In other words, the sample would differ in predictable ways given its location. This type of reasoning is a by-product of responding to one of the planning prompts.

5.2.6. Frequency-based reasoning

Frequency-based reasoning involves grouping responses to an open-ended survey question into categories on the basis of frequency, which reduces the number of categories. This type of reasoning was evident when Group A discussed the possibility of grouping infrequent responses into a general category called "other." The results of the group's survey on pets were 9 dogs, 5 fish, 4 cats, 4 birds, 3 turtles, 1 lizard, and 1 gerbil. Dana and Joly disagreed about whether or not infrequent pets such as gerbil, turtle, and lizard should be grouped into a category "other." The dialogue below illustrates Dana's frequency-based reasoning.

Dana: Yeah, but then, 'cause look. The only other animals there are, are gerbils, lizard, and
Joly: Lizards, turtles.
Dana: and turtles. So if we do turtles then we're only gonna have two and then there's no point in having an other.
Joly: I know. What's the point in having an "other"?
Dana: Because, in case somebody doesn't have those ones.

5.2.7. Organization-based reasoning

Organization-based reasoning consisted of explaining how to enter or organize nominal data into a spreadsheet or how to group responses or categorize different types of responses they obtained from their open-ended survey into a spreadsheet or a graph. An example of this type of reasoning is an exchange between Cathy and Jamie (Group B) that occurs as a result of Cathy's concern about the sample size (40 students, 10 per grade), which she thinks is too many. Cathy exclaims "If you have 40 kids then how are you gonna make a graph with 40 individual things? You have all different things." Jamie responded, "You don't! You put them together into a bar graph or a line graph." The issue here is how to organize the types of professions respondents gave. Cathy thinks that each individual's response has to be represented separately, whereas Jamie understands that each response can be organized into types or groups of responses. Joly also demonstrated organization-based reasoning when she suggested that the data be organized by rank order.

5.2.8. Interpretation-based reasoning

Interpretation-based reasoning is evident when students explain measures of central tendency (or averages), reason about their choice of measure or graph, explain why a graph does not work, or interpret data and draw conclusions from a graph. Students' decisions are made to enhance the interpretation of the data or to answer the question appropriately. For example, Joly (Group A) explains that the group will "use pie charts statistics because I think when most people glance at it they realize which piece is bigger." In other words, the graph allows her to quickly interpret the data. In addition, Dana argues for the mode because "we want to know what number of pets are the most." A final example is seen when the teacher asked Group B to explain why they obtained different responses from the different grade levels. Jamie replied, "Maybe the grade 11s ... have ... a more believable approach. Like you want to go to university, and like be a businessman instead of just like being a model, which you can but it's more unlikely."

5.2.9. Alignment to question reasoning

This type of reasoning applies to situations in which students reason that they must make decisions that are consistent with the survey question. Group A's discussion during data organization revealed that students were confused about which question they were answering because respondents were asked about the number and types of pets but the group's initial focus of the survey was on the number of pets. When the group discussed data organization Dana kept saying that the group had to stick with the question and would reject Joly's suggestions because "that was not our question!"

5.2.10. Standardization oriented reasoning

Standardization oriented reasoning refers to selecting the same number of categories or respondents in each comparison group to allow for meaningful interpretation. If standardization in this way is not possible then it can be accomplished by converting the data in some way. For example, Cathy from Group B was concerned about asking a large number of people to respond to a free-response survey saying, "But everybody wants to be different things. It doesn't make sense. You have to understand." In response Jamie explained, "It doesn't matter how many people we have [because] it's just percentages. It's not one or two people, it's one percent." Jamie's comment reflects his solution—i.e., to standardize by converting frequencies into percentages." Later, Cathy emphasized sampling an equal number of students per comparison group suggesting a sample size of 40 because "you can't divide 30 up into these groups! It doesn't fit." Jamie affirms her idea: "Let's do 40. It kind of fits well. 1, 5, 9, 11 [grades]."

5.3. The link between modeling activities, modes of statistical reasoning, and teacher scaffolding during inquiry

Most of the statistical reasoning modes can be traced to the learning activities. In all of the activities (i.e., the introduction, pulse rate activity, and LE activity) the teacher emphasized making predictions from questions posed. This process makes students sensitive to the responses that they should expect to a question. Thus, although considering the range of a population's interests was not specifically addressed in all activities, the value of anticipating responses was clearly communicated and could have provided a basis for variety-based reasoning. Specifically, the discussion of the "other" category could have been a

springboard for thinking of atypical responses and having a way to identify them. In other words, this category allows for a range of responses.

The LE activity and discussion of the "other category" also provided the basis for two types of reasoning: frequency-based reasoning and characteristics oriented reasoning. Dana (Group A) often referred to the LE example to justify organizing infrequent responses into an "other" category. Presumably, the less typical responses that one would expect to be in the "other" category for an open-ended question are also the less frequent. This type of thinking reflects frequency-based reasoning. The LE also modeled characteristics oriented reasoning. One exemplar discussed in class consisted of a student who explained that the results of the favorite restaurant survey would be different if it had been conducted in another province given the different demographics and interests.

The organization and interpretation of data were modeled throughout the three learning activities. For example, Cathy's concern about the sample size and its impact on the graph can be traced back to the pulse rate activity where students constructed graphs representing the mean and individual responses. Cathy thought that the 40 responses had to be presented individually, whereas Jamie realized that the data could be grouped to present an average (the mode) as was also done in the pulse rate activity. These are instances of organization-based reasoning. Interpretation based reasoning was always encouraged as the teacher asked groups to make a decision about a question or a graph and to present a compelling case for that decision. Graph selection was discussed in the context of facilitating interpretation as well as when graphs were preferable to tables in communicating information, such as in the pulse rate activity.

Four modes of statistical reasoning cannot be traced directly to the introduction, pulse rate activity, or the LE activity: standardization-based, reasoning about alignment, LLN oriented, and category level. Prior knowledge and experience (e.g., a plausible precursor to standardization-based reasoning is students' previous mathematics experience), as well as the specific inquiry experienced by the groups likely played a role. For example, reasoning about the alignment to the question was necessarily context-specific in this study as it was based on students coming to terms with what the group consensus was on the research question. Nonetheless, alignment is critical in inquiry. Consequently, this mode of reasoning could be evident in other studies of inquiry in statistics or science.

The teacher did check in on the groups at various points during their inquiry. He clarified the planning prompts to students when they did not remember what a particular concept meant, he urged them to think of a question worthy of investigation, he suggested ways of collecting data that would be practical (e.g., advisor groups), and he assisted students when they encountered computer problems. Thus, the teacher's scaffolding was critical to the groups' design of an investigation and in helping them understand statistical ideas. Moreover, the teacher's guidance in selecting a survey question may partly explain the high incidence of reasoning during problem posing. However, the teacher's intervention did not explain the reasoning documented in the case analysis. In other words, students' reasoning was more closely related to the teacher's modeling prior to the investigation than it was to his scaffolding during the inquiry process.

6. Discussion

The purpose of the case study was to document the modes of statistical reasoning that are possible when two groups of grade 7 students are given opportunities to engage in all

phases of a statistical inquiry. We provided simple models of inquiry that engaged students in specific learning activities to give them a framework for thinking about statistical content that was relevant. Rather than expecting highly sophisticated forms of reasoning the intent was to provide situations for students to engage in statistical thinking and to document whatever modes of statistical reasoning emerged. We found that the students reasoned in a variety of ways that reflected both their personal knowledge and the framework. Moreover, students' statistical reasoning reflected primitive forms of thinking that are emergent and thus have the possibility of growth (Shaughnessy, 1992). The fact that we observed such reasoning in a short study shows promise for starting with simple models that can be expanded through instruction and inquiry.

In our study we identified ten modes of statistical reasoning, several of which can be viewed as underlying the types of statistical reasoning that Garfield and Gal (1999) identified as key goals for instruction. We have organized our discussion of the statistical reasoning modes around the statistical reasoning types and we refer to previous studies to contextualize the significance of the findings where relevant. We also discuss modes of statistical reasoning that do not relate directly to Garfield and Gal's (1999) statistical reasoning types, which we suggest calls for an elaboration of the statistical reasoning framework.

6.1. Reasoning about samples

Key ideas underlying the ability to reason about samples include the relationship between a sample and a population and the representativeness of a sample to its population (Garfield & Gal, 1999). Understanding these issues affects what inferences people can make about a population on the basis of a sample. In our case study, we found two reasoning modes—LLN oriented and characteristics oriented—that reflected some signs of representative thinking (albeit incomplete)—i.e., thinking about how to sample individuals from a population (e.g., a large and carefully selected sample is more representative of the population than a small and biased sample) and making inferences about findings given the sample characteristics and size.

6.1.1. LLN oriented reasoning

LLN oriented reasoning is a precursor to being able to think about how the size of the sample affects the variability within the sample, which in turn impacts the representativeness of the sample to the population, and thus the validity of the results. LLN oriented reasoning was evident when students explained the need for a large sample in order to obtain accurate results. We would hope to see this mode of reasoning during data analysis or representation when students interpret and attempt to generalize their findings, as well as during data collection. Indeed we found instances of LLN oriented reasoning in all inquiry phases. This reasoning appears to be based on students' prior knowledge since the teacher did not explicitly address the role of sample size in sampling and drawing conclusions.

The incidence of LLN oriented reasoning in our study supports previous studies with middle school students (Watson & Moritz, 2000a). However, our students were not always consistent in reasoning this way. For example, two students within a group reasoned according to the LLN principle while a third group member reasoned about sampling decisions based on task constraints. Moreover, students' grasp of sampling was fragile. Although there was some LLN oriented reasoning, most students did not articulate their

reasoning about sampling method. Moreover, the sampling methods they employed in their surveys, such as sampling from a self-generated list of respondents and from diverse groups (e.g., advisor groups), were not normative. Nonetheless, they could be precursors to developing an understanding of random sampling.

Watson and Moritz (2000b) found that 6th grade students grasped either sample size (i.e., by suggesting a large sample) or sampling method (i.e., by suggesting stratified sampling) but not both. These findings suggest that an understanding of sampling may be piecemeal in the early middle school years and that achieving a complete grasp of sampling may be progressive (Watson & Moritz, 2000b).

6.1.2. Characteristics oriented reasoning

This mode of reasoning occurred in our study when groups considered sample characteristics in interpreting their findings in the data analysis or representation phases. It is an example of "reasoning about samples" in the sense that the students are thinking about "how samples are related to the population and what may be inferred from the sample" (Garfield & Gal, 1999, p. 211). Students thought about how the results would change with a new sample whose characteristics differ from the current sample to explain how applicable the findings were to the population. The students did think about sample representativeness, but not exactly in the way that one would expect, that is, in terms of the representativeness of the existing sample.

Students' characteristics oriented reasoning occurred in response to a specific planning prompt for data analysis (i.e., would your results change if the study were done differently and why?), which explains why this reasoning mode occurred only in this phase. Perhaps, students did not reason normatively about sample characteristics when they responded to the prompt because they lacked sufficient knowledge of sampling method, which requires thinking about sample characteristics. Nonetheless, providing students with an opportunity to think about the representativeness of the sample in terms of its characteristics through the data analysis prompt could serve as a bootstrap for building students' thinking of representativeness in the context of sampling method.

6.2. Reasoning about data

According to Garfield and Gal (1999), reasoning about data involves "recognizing or categorizing data as quantitative or qualitative, discrete or continuous, and knowing how the type of data leads to a particular type of table, graph, or statistical measure" (p. 210). Three modes of reasoning from our study are relevant to this reasoning type: organization-based reasoning, category-level reasoning, and frequency-based reasoning.

6.2.1. Organization-based reasoning

The capabilities underlying reasoning about data enable students to reason about how the data should be structured in a spreadsheet to draw conclusions from the resulting graph. For example, setting up a spreadsheet in which the cells to be computed contain frequencies associated with nominal categories suggests an understanding that the categories are qualitative and that to perform any computations or create graphs one must organize the categorical data into frequencies. In contrast, not grouping individual responses hampers the interpretation of data. The reasoning mode in our study that

involves explaining how to enter or categorize (or not categorize) the data in a spreadsheet or graph is organization-based reasoning, which we found in the data analysis and representation inquiry phases. Such reasoning can be traced to the pulse rate activity in which graphs were constructed using individual data and grouped data (e.g., mean).

In previous work, organizing data to represent each individual data point with its own bar on a graph was referred to as individual-based thinking while grouping data into averages, each represented by a bar, was referred to as aggregate-based thinking (Bright & Friel, 1998; Hancock et al., 1992). We documented both types of thinking in our case analysis (except that our examples involved grouping into categories rather than summaries), and chose to use the term organization-based as a general term that encompasses individual- and aggregate-based reasoning.

6.2.2. Category-level reasoning

When inquiry involves surveys an added element to consider is how to organize the responses gathered from a free-response survey into meaningful categories containing options that are similar in kind. This means that students sometimes have to think about which concepts are more hierarchical than others (e.g., entertainment vs. actors and singers) and decide which level of the hierarchy to use in grouping the data, in addition to which concepts belong together. This was the case in our study. Category-level reasoning involved thinking about the representative features of a category with respect to another category while classifying data. This mode of reasoning is consistent with reasoning about data in that it underlies the ability to recognize or group data as either quantitative or qualitative. In this case, students determined that the data were qualitative, but part of that thinking also involves deciding how to organize the categories in a way that accurately represents the data. Thus, students were reasoning about data but about issues that underlie the main ideas articulated in Garfield and Gal's (1999) definition.

Category-level reasoning reflects ideas from concept formation studies in which children show a preference for basic level concepts (Murphy, 2002). The issue of what constitutes a typical pet, which was raised in one group's discussion of a category "other" during data analysis, relates to basic level concepts. One student argued for grouping uncommon pets into this "other" category while another countered that the label was uninformative and did not represent pets. The implication of the former student's point is that infrequently occurring objects are atypical and thus not prototypical basic level concepts. By rejecting the "other" category, the second student implied that infrequent pets were still members of the pet category regardless of their typicality, and thus could be subsumed within a general category.

Category-level reasoning has not been previously reported in the inquiry literature. Perhaps, this is because the way of classifying data in this reasoning mode is more typical of surveys than experiments given the nature of the data and because few studies have focused on survey inquiry. The LE exemplar in which five categories are used (i.e., types of fastfood restaurants) with a category called "other" to accommodate other tastes could also explain category-level (and frequency-based) reasoning. Though the actual thinking reflected in these reasoning modes was not modeled in the learning activities, the example may have provided the impetus for their occurrence. It would be worth further examining the link between categorization and reasoning about data in future work.

6.2.3. Frequency-based reasoning

In frequency-based reasoning responses were grouped into categories based on the frequency with which they occurred in the survey rather than on typicality. Infrequent responses were grouped into a general category (e.g., other) for this reason and to limit the number of categories to a reasonable number. As with category-level reasoning, this mode of reasoning involves reasoning about data but reflects a specific way of thinking about classifying categorical data. Frequency-based reasoning was less common in our data and led to lower rater agreement. Future research may determine the significance of this category.

6.3. Reasoning about statistical measures

Capabilities that underlie students' reasoning about statistical measures include the following: knowing that measures of center and spread provide valuable information about data sets, which can be compared; that both measures of center and spread must be used in conjunction to make accurate inferences about data sets; and that to do so, students must be able to determine which measures are appropriate and under which conditions (Garfield & Gal, 1999). In addition, reasoning about measures involves making inferences about summaries of data based on the sample size (i.e., summaries derived from large samples are more accurate than those from small samples).

Two modes of reasoning in our study involve reasoning about statistical measures: interpretation-based and standardization oriented. Interpretation-based reasoning involves choosing measures and graphs, interpreting data represented in either form, and drawing conclusions from them. Typically, reasoning in this way occurs in the data analysis and representation phases of inquiry. It is consistent with reasoning about measures in that it involves being able to make decisions about measures of center that enable meaningful interpretation of data.

Standardization oriented reasoning involves reasoning for categories in the question to make a meaningful comparison, for the same number of respondents in each comparison group, and for converting frequency data to make meaningful comparisons. It was therefore evident in the problem posing, data collection, and data analysis inquiry phases. Standardization oriented reasoning can be viewed as providing a foundation for the ability to reason about data sets that are being compared, which is one capability that underlies reasoning about measures. In this sense, the reasoning mode is a part of the reasoning type. Standardization oriented reasoning was not common in our study, which may reflect the fact that it is a minor consideration or that standardization was not a major issue in this particular context.

6.4. Reasoning about data representations

This type of reasoning involves the ability to comprehend information in graphs (i.e., read and interpret), to construct and modify graphs in a way that best represents the sample and is appropriate to the data being displayed, and to determine the shape, center, and spread of a distribution of data (Garfield & Gal, 1999). Interpretation-based reasoning is consistent with reasoning about data representations in cases where students make a choice about the type of graphical display to represent and interpret their data.

6.5. Reasoning about uncertainty

Key ideas that must be understood to reason successfully about uncertainty consist of randomness, chance, and uncertainty (Garfield & Gal, 1999). Using these ideas to ascertain the likelihood of an outcome is a fundamental activity in this type of reasoning. The framework that we provided and the type of inquiry that students engaged in did not introduce or elicit activities pertaining to this type of reasoning. Consequently, no modes of statistical reasoning that underlie reasoning about uncertainty were found in this study.

6.6. Reasoning about association

This type of statistical reasoning occurs when two variables and their relationship are examined. Consequently, students must be able to interpret the relationship, interpret a two-way table or a scatterplot depicting the association, and grasp the distinction between correlation and causation (Garfield & Gal, 1999). Again, since neither the framework nor students' inquiries involved more than one variable, we did not observe modes of reasoning that could underlie reasoning about association.

6.7. The missing reasoning type: reasoning about questions

Garfield and Gal's (1999) reasoning types focus on critical issues underlying statistical reasoning and encapsulate much of the thinking that underlies inquiry in this context. The only inquiry aspect not captured in their framework involves reasoning about research or survey questions. A fair amount of our students' reasoning centered on the survey question, in particular, the selection of categories for the question. While one might argue that the reasoning modes we identified pertaining to survey questions were idiosyncratic to this study, keep in mind that statistics inquiry in classrooms often begins with surveys and then progresses on to experiments (NCTM, 2000). Moreover, a discussion of "what's a good or researchable question" is relevant to all types of inquiry. Three statistical reasoning modes in our study focused on the general issue of selecting categories for a survey question, which could be one characteristic of reasoning about questions: population relevant, variety based, and category-level.

6.7.1. Population relevant reasoning

In survey situations, the concept of representativeness underlies the thinking that takes place when one considers how to sample categories that reflect the characteristics of the population. Population relevant reasoning in our study consisted of students reasoning about the relevance of survey categories based on perceived interests of the population. There is evidence that this mode of reasoning does occur during problem posing. In Hancock et al.'s (1992) study students rejected a certain type of music because they thought that no one actually listened to it. This way of reasoning was also found in studies involving the sampling of individuals. Schwartz and Goldman (1996) suggested that statistical rules for sampling are less evident when individuals as opposed to neutral objects, such as M&Ms, form a population because both

characteristics (e.g., age, gender, political affiliation) and opinions/interests of that population are sampled. Students in our study focused on population interests in sampling categories by anticipating responses to a survey question based on their knowledge of the population. Emphasis on population interests rather than population characteristics reflects a misapplication of statistical ideas. The case analysis suggests that the influence of personal knowledge on reasoning is not limited to sampling situations but extends to the sampling of categories (i.e., types of professions) during problem posing.

6.7.2. Variety-based reasoning

Variety-based reasoning refers to representing a variety of the population's preferences in survey categories. This reasoning mode focused on a range of categories to represent diverse interests in the population. The notion of representativeness is not the focal point in variety-based reasoning because the emphasis is on the number of categories. However, the categories must still be representative of the population. Thus, representativeness lies in the background of thinking about a range of categories, but it is still part of that reasoning. In essence, population relevant and variety-based reasoning each are a different side of the same coin.

Until now, variety-based reasoning has not been documented with respect to problem posing. However, there is evidence of such reasoning when students think about sampling (Jacobs, 1999). In judging the validity of various sampling techniques, the middle school students in Jacobs's (1999) study stated that it was unsound to sample groups of people who were more likely to select a particular response because doing so would result in the representation of a single opinion. Hence, a range of groups or individuals with different opinions must be sampled. The focus on opinions is reminiscent of population relevant reasoning. However, the emphasis on representing a range of opinions is consistent with variety-based reasoning. In this case, students want a survey topic that will generate different responses or they want to sample a sufficient number of categories to ensure that multiple preferences are represented.

According to Jacobs (1999), variety-based reasoning is based on the notion of fairness (i.e., it is unfair to represent one point of view), which provides a foundation for developing an understanding of randomness. Students in our study may have been sensitive to this issue as a result of a whole class discussion during the LE activity dealing with a category "other." Students thought that including this category was important because it allowed for all of the opinions to be reflected in the question. We wonder whether this notion of fairness is the reason why students in both groups insisted on sampling genders equally. Since students did not discuss or explain why they planned to sample in this way, we could not code for reasoning. However, the parallel is intriguing. Future work could explore the possibility that variety-based reasoning is the basis for sampling each gender in equal numbers.

6.7.3. Category-level reasoning

This statistical reasoning mode was discussed in relation to the "reasoning about data" type because it focused on the classification of data for analysis and graphical representation. This mode can also occur during problem posing when students focus on the generality or specificity (or both) of categories for a survey question. One student

in our study preferred to add subordinate level categories (e.g., astronaut) associated with basic level concepts (e.g., scientist) during the problem-posing phase because respondents might not interpret the categories in the same way. The selection of categories can make a difference in the validity of the responses obtained, and deciding which categories to use and whether they mean the same thing to everyone is an important issue.

6.8. The statistical reasoning mode without a home: alignment to question reasoning

Garfield and Gal (1999) identified types of statistical reasoning that targeted mental activities on statistical tasks in general, not inquiry in particular. Thus, while many of the types they point to apply to inquiry, there are aspects of reasoning in this context that are not addressed. As mentioned previously, one of these is the process of generating questions and survey categories. Another involves reasoning about all inquiry phases based on the question driving the inquiry, such as that found in our study. Reasoning in terms of alignment to the question involves making decisions that are consistent with the question being investigated. It arose in our study because of specific challenges faced by a group who had collected more than one type of data for its survey. This action led to confusion when the data were organized to answer the question. Relating the data to the question when that question was unclear resulted in reasoning in terms of alignment.

Although this mode of reasoning seems particular to the circumstance, one would want students to be able to keep the question in mind at all inquiry phases; after all, the question does drive the inquiry. Moreover, revising a question based on additional data gathered does happen in inquiry and beginning students can be uncomfortable with this fact. Perhaps, reasoning in terms of alignment should be a particular type of statistical reasoning during inquiry.

7. Conclusion

Students in this study did grasp the inquiry process and several of the concepts related to the phases of inquiry. This understanding was based on previous knowledge, the introduction and modeling activities, and the teacher's scaffolding during the group inquiry task. We found that the students reasoned in several ways or modes during problem posing (i.e., population relevant, variety-based, category-level, standardization oriented reasoning), data collection (i.e., LLN oriented, characteristics, standardization oriented), and data analysis/representation (i.e., alignment to question, standardization oriented, frequency-based, category-level, organization-based, and interpretive). Most statistical reasoning modes are contextualized particulars of the statistical reasoning types Garfield and Gal (1999) identified as goals for instruction, namely, reasoning about sampling, reasoning about data, reasoning about statistical measures, and reasoning about data representation. The modes we documented capture some but not all aspects of thinking underlying the statistical reasoning types, partly because our framework exposed students to statistics in a short time frame.

The statistical reasoning we found tended to be naïve; however, this provides a good foundation for more sophisticated ways of reasoning that can be fostered in other inquiry

experiences. There is evidence for the validity of certain modes of statistical reasoning that emerged from this study as they have been reported in previous studies, albeit in different inquiry phases. This added information is noteworthy because it provides a fuller picture of statistical reasoning across inquiry phases.

Reasoning about research or survey questions was not identified by Garfield and Gal (1999) as a type of statistical reasoning, and yet modes of reasoning relating to this type were the second most prevalent in our study. This finding suggests that we consider expanding the statistical reasoning framework. Moreover, the number of different ways that students can think about a survey question highlights the potential complexity in children's thinking in this situation, and the value of using survey investigations as a platform to learn about their thinking.

8. Limitations and future directions

This case study is limited in three ways. First, the possibility that the students could conduct an experiment or a survey was not clearly articulated to them. Even though the teacher modeled an experiment with the pulse rate activity, it is possible that the recency of the library of exemplars and the time spent discussing the criteria in the library led students to restrict their inquiries to surveys. Second, the exemplars of surveys for each criterion focused on a specific aspect of the criterion and were not sufficiently rich. The teacher's introduction provided other examples of surveys, which was helpful. However, additional examples of different types of inquiry should be included in the library of exemplars. As found in previous work, people learn from models but they need to be sufficiently rich (Lajoie et al., 1998). Finally, as this was a case study, we need to determine whether we would find the same or similar modes of reasoning with a larger sample.

It is important to note that the findings in this study pertain to survey situations. Thus, the reasoning modes identified during problem posing may be less evident during other types of inquiry, such as experiments. Studies that examine reasoning during problem posing in various types of inquiry are needed to confirm this possibility. In addition, we suspect that surveys may increase the salience of and reliance on personal knowledge. Future research could examine the role that various types of knowledge (e.g., real-world, statistical, scientific) and beliefs have on reasoning as students pose problems involving variables that are tied to substantive content.

From an instructional standpoint, starting with surveys is a good first step in learning inquiry because it can build on students' personal knowledge and provide a basis for scientific inquiry. Given that learners' personal interests drive their inquiries we need to assist them to better integrate these interests with statistical content in conducting surveys. This integration would lead to growth of statistical reasoning in a planful manner. Future work can build on these research findings by examining whether specific scaffolding can lead to more sophisticated statistical knowledge in this age group.

Acknowledgments

We would like to acknowledge Bill Nevin for his support and collaboration in collecting this data, as well as Luis Marquéz and Jésus Carlos Guzman from the University of

Mexico and the Applied Cognitive Science research group at McGill University for assistance in videotaping and transcribing. Finally, we are grateful to Ralph Ferretti for valuable comments on this manuscript.

Appendix A. Statistical thinking conveyed in the introduction and modeled in each instructional activity by inquiry phase

A.1. Question posing

General introduction

- *Ideas generated*: How many people in Canada are smokers and nonsmokers? How far do professional golfers hit the ball in a tournament?
- *Newspaper articles*: In your opinion does the referendum question as worded express clearly that Québec wants to be a sovereign country? Is it better to build a new baseball stadium that is outdoors or to keep the old indoor stadium? These questions are asked to make a case for a particular point of view.
- *Population*: People in Québec.

Pulse rate activity

- Are the at-rest pulse rates of grade 7 students different from their pulse rates after physical exercise?
- We usually have a prediction of what the answer will be when we ask a question (refers to the baseball stadium example).
- *Prediction*: Pulse rates after exercise will be higher than pulse rates before exercise.
- *Population*: Defined as whom we are interested in—i.e., grade 7 students.
- *Variable*: Defined as the thing that varies or that changes. It is what is measured—i.e., pulse rate.

Library of exemplars activity

- In the pulse rate activity it would not have been sensible to have a general question such as "What happened to the grade 7 pulse rates?" We needed a specific and clear question like the one that was used. The "appropriate" clip shows a general question "What is your favorite fastfood restaurant?" The "more appropriate" clip shows that more a specific question with categories had been asked: "What is your favorite fastfood restaurant between McDonald's, Harveys, Burger King, Lafleur's, other?"
- *Variable*: The thing that the class is thinking about: favorite fastfood restaurant.
- *Categories*: McDonald's, Harveys, Wendy's, Burger King, Lafleur's, other.
- After randomly selecting different groups to present their argument, the consensus was that the question with the categories was thought to be best because: (a) it gives options, which makes it clear; (b) it has a category "other" that covers all the bases in case the preferred option is not in the list of categories (otherwise the survey would yield an inaccurate response); and (c) the general question can yield 20 completely different restaurants so options should be limited.

A.2. Data collection

General introduction

- *Sample*: The pollsters did not ask everyone in Québec they asked a certain number of people.
- *Representativeness*: They also did not ask the first 100 houses in the most Anglophone part of the city; must ask all kinds of people.

Pulse rate activity

- *Sample*: Cannot ask every grade 7 student in Québec but can ask grade 7 students in this class since the data must be collected tomorrow.
- *Representativeness*: The class is not a very good sample. It would be better to go to different places (location) and ask people with different interests. However, the class is the sample that is available.
- Everyone takes their pulse rate twice: once at rest and once after jogging on the spot for 5 min in the hallway. Each time they write their pulse rate on a small piece of paper.

Library of exemplars activity

- *Sample size*: How many people you ask, how much data you collect. A student suggested sampling 50 people in a few places. (clips did not work so teacher built on the research question examples).
- *Group size*: How many people are considered in all [not in the guidelines].
- *Sample*: People who eat fastfood—kids because they really like fastfood.
- *Location*: The mall because a lot of kids hang out there or a school.
- *Representativeness*: Is the school typical—is it the same as others? In response a student suggests asking a couple of children from each grade. Teacher supports this idea.

A.3. Data analysis

General introduction

- Not addressed separate from graphs.

Pulse rate activity

- Data for each student is entered into a spreadsheet in 2 columns: Before, after. Students eyeball the data and infer that most pulse rates go up.
- *Outlier*: Introduced when a student points to an abnormally low at-rest pulse rate. Defined as a piece of data that does not fit with everything else.
- *Averages* can be used to estimate of how much the pulse rates have gone up.
- *Mean*: Add up the numbers and divide by number of entries. Calculated using software: Before = 37, After = 63.
- *Median*: Analogy of median in the road—the middle. Before median = 37, After median = 60. The data were sorted using the software to find the median.

- *Mode*: Defined as the one that happens the most. Results are bi-modal. Teacher says that the mode is not very good.
- Students must convince classmates whether the mean or median is most compelling. Teacher concludes that both types of averages are good, and that the average pulse rates go up by 25.

Library of exemplars activity

- The mode, median, and mean are revisited as a way to analyze data.
- *Range*: Defined as the difference between the largest and the smallest.
- The appropriate clip shows an output showing the number of cases, the minimum value, maximum value, and the range. No interpretations of this table are made in the clip. The more appropriate clip shows a student explaining that the results would be different if the study had been conducted in another province given the different demographics and thus different interests. In this clip the results are interpreted.
- There is not much discussion of these exemplars, other than the teacher informing students that they will have to consider these aspects of data analysis in their inquiry projects.

A.4. Data representation

General introduction

- Tables and graphs show the answer to questions.
- A table and a bar graph show results of different pollsters in percentages. The graph is better because the size of the differences can be quickly ascertained without looking at the numbers while tables require more work.
- A bar graph compares the average number of spectators who attended football games in the old stadium and the average of those who attended games in the new stadium over a period of 1 year. The graph was presented to make a case for building a new outdoor baseball stadium. The large difference was emphasized by the contrasting colors for the bars on the graph.

Pulse rate activity

- Which data should be used in a graph to show difference: raw data or averages? Both are used for illustrative purposes.
- Groups are randomly selected to present their argument for best graph.
- *Bar chart of individual data points*: Students indicate that this graph presents the same information as the stacked graph but in a different way.
- *Bar graph showing averages*: Students conclude that the after rates are higher than before rates.
- *Stacked bar graph showing individual data points*: Students have difficulty interpreting the graph. The teacher points out that it is hard to tell that the average has increased. A student asks why the bars go up to 600 when no one had that pulse rate. Some find graph useful because the cumulative rates still show an increase.

- *Line graph of individual data points*: Shows each data point including outliers but difficult to make sense of.
- *Line graph with sorted individual data points*: More organized, easier to see the trend. The graph may be a little dishonest because it hides the outlier. Considered the best one.
- *Connection to research question*: Is the research question "how the pulse rate increased for each person?" No it is to find out overall whether the rates change from before to after.

Library of exemplars activity

- Appropriate clip shows a graph without explanation. The more appropriate clip shows four tiled pie charts to compare the results of each group member.
- *The teacher's criteria for good a graph*: Color for contrasting differences, interesting, and easy to interpret.
- Advantage of graphs over tables: Immediate interpretation.

References

American Statistical Association. (1991). *Guidelines for the teaching of statistics K-12 mathematics curriculum*. Landover, MD: Corporate Press.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bracewell, R. J., & Breuleux, A. (1994). Substance and romance in the analysis of think-aloud protocols. In P. Smargorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 55–88). Newbury Park, CA: Sage.

Bright, G. W., & Friel, S. N. (1998). Graphical representations: Helping students interpret data. In S. P. Lajoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12* (pp. 63–88). Mahwah, NJ: Erlbaum.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal for the Learning Sciences, 6*(3), 271–316.

Claris Corporation. (1992). ClarisWorks 4.0 [Computer program]. Santa Clara, CA: Apple Computer Inc.

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

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.

English, L. D. (1998). Children's problem posing within formal and informal contexts. *Journal for Research in Mathematics Education, 29*(1), 83–106.

Evans, J. St. B. T. (1993). The cognitive psychology of reasoning: An introduction. *Quarterly Journal of Experimental Psychology, 46A*, 561–567.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children (C.A. Sherrard, Trasn.)*. Boston: D. Reidel.

Fredriksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1 and 2), 99–125.

Garfield, J., Gal, I. (1999). Teaching and assessing statistical reasoning. In L. V. Stiff, F. R. Curcio (Eds.), *Developing mathematical reasoning in grades K-12. 1999 Yearbook* (pp. 207–219). Reston, VA: National Council of Teachers of Mathematics.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*, 337–364.

Holyoak, K. J., & Morrison, R. G. (2005). Thinking and reasoning: A reader's guide. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 1–9). New York, NY: Cambridge University Press.

Jacobs, V. (1993). Stochastics in middle school: An exploration of students' informal knowledge. Unpublished master's thesis, University of Wisconsin, Madison, WI.

Jacobs, V. (1999). How do students think about statistical sampling before instruction? *Mathematics Teaching in the Middle School, 5*(4), 240–263.

Jones, G., Thornton, C., Langrall, C., & Mooney, E. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*(4), 269–307.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*(2), 123–141.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 24*(5), 392–414.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing R., et al. (2002). *Students' use of modal clumps to summarize data*. Paper presented at the 6th meeting of the International Conference on Teaching Statistics (ICOTS6), Cape Town, South Africa.

Lajoie, S. P. (1997). Technologies for assessing and extending statistical learning. In J. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 179–190). Amsterdam, the Netherlands: IOS Press.

Lajoie, S. P. (1999). Understanding of statistics. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 109–132). Mahwah, NJ: Erlbaum.

Lajoie, S. P., Lavigne, N. C., Munsie, S. D., & Wilkie, T. V. (1998). Monitoring student progress in statistics. In S. P. Lajoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12* (pp. 199–231). Mahwah, NJ: Erlbaum.

Lavigne, N. C., & Lajoie, S. P. (1996). Communicating performance standards to students through technology. *Mathematics Teacher, 89*(1), 66–69.

Lehrer, R., & Schauble, L. (2000). Inventing data structures for representational purposes: Elementary grade students' classification models. *Mathematical Thinking and Learning, 2*(1 and 2), 51–74.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635–679.

Leighton, J. P. (2004). Defining and describing reason. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 3–11). New York, NY: Cambridge University Press.

Metz, K. E. (1998). Emergent ideas of chance and probability in primary grade children. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 149–174). Mahwah, NJ: Erlbaum.

Murphy, G. L. (Ed.). (2002). *The big book of concepts*. Cambridge, MA: A Bradford Book, The MIT Press.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Nisbett, R. E. (Ed.). (1993) *Rules for reasoning* (pp. 297–314). Hillsdale, NJ: Erlbaum.

Petrosino, A. J. (1998). *The use of reflection and revision in hands-on experimental activities by at-risk children*. Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN.

Piaget, J., Inhelder, B. (1975). *The origin of the idea of chance in children*. (L. Leake, Jr., P. Burrell, & H.D. Fischbein, Trans.). New York: W.W. Norton. (Original work published in 1951).

Russell, S. J. (1999). Mathematical reasoning in the middle grades. In L. V. Stiff & F. R. Curcio (Eds.), *Developing mathematical reasoning in grades K-12* (pp. 1–12). Reston, VA: National Council of Teachers of Mathematics.

Schwartz, D. L., & Goldman, S. R. (1996). Why people are not like marbles in an urn: An effect of context on statistical reasoning. *Applied Cognitive Psychology, 10*, 99–112.

Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & CTGV (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12* (pp. 233–273). Mahwah, NJ: Erlbaum.

Silver, E. A., Mamona-Downs, J., Leung, S. S., & Kenney, P. A. (1996). Posing mathematical problems: An exploration study. *Journal for Research in Mathematics Education, 27*(3), 293–309.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York: National Council of Teachers of Mathematics and Macmillan.

Watson, J. M., & Moritz, J. B. (2000a). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*(1 and 2), 11–50.

Watson, J. M., & Moritz, J. B. (2000b). Developing concepts of sampling [Electronic version]. *Journal for Research in Mathematics Education, 31*(1), 44–70.

Watson, J. M., & Moritz, J. B. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments [Electronic version]. *Journal for Research in Mathematics Education, 34*(4), 270–304.

Zimmerman, C. (2000). The development of scientific reasoning skills [Electronic version]. *Developmental Review, 20*, 99–149.

Zimmerman, C. (2005). *The development of scientific reasoning: What psychologists contribute to an understanding of elementary science learning*. Paper commissioned by the National Academies of Science (National Research Council's Board of Science Education, Consensus Study on Learning Science, Kindergarten through Eighth Grade). Final report available from http://www7.nationalacademies.org/bose/Corinne_Zimmerman_Final_Paper.pdf.