High School Students' Levels of Thinking in Regard to Statistical Study Design

Randall Groth
Salisbury University
<regroth@salisbury.edu>

The study describes levels of thinking in regard to the design of statistical studies. Clinical interviews were conducted with 15 students. Each student was enrolled in high school or was a recent graduate. The students interviewed represented a range of mathematical backgrounds. During the clinical interview sessions, students were asked how they would go about designing studies to answer several different quantifiable questions. Several levels of sophistication were identified in their responses. The levels of sophistication in response are discussed in terms of the Biggs and Collis (1982, 1991) cognitive model.

Study design is foundational to the practice of statistics. Cobb and Moore (1997) underscored this point with the following statement:

> Statistical ideas for producing data to answer specific questions are the most influential contributions of statistics to human knowledge. Badly designed data production is the most common serious flaw in statistical studies. Well designed data production allows us to apply standard methods of analysis and reach clear conclusions (p. 807).

Wild and Pfannkuch (1999) also emphasized the importance of study design, stating that it is an indispensable part of the overall process of statistical thinking. Recognizing the importance of study design, the National Council of Teachers of Mathematics (NCTM) (2000) recommended that students should begin to have experiences in designing simple studies during their preschool years and develop increasingly more sophisticated study design strategies throughout their years of formal schooling.

Purpose of the Study

Since the topic of study design forms an important part of statistics education, the need exists to understand students' patterns of thinking in response to statistical study design tasks. Research that describes students' cognition in regard to mathematical topics has the potential to help improve the teaching of the topics (Fennema & Franke, 1992; Even & Tirosh, 2002). The purpose of the present study was to contribute to the knowledge base concerning students' understanding of study design by addressing the following two research questions:

i. What are the defining characteristics of high school students' (high school in the U.S., where the present study was conducted, generally includes students of 14-18 years of age) patterns of response to statistical study design tasks?
ii. What cognitive level can be associated with each of the patterns of response identified?

Previous Research on Students' Knowledge of Study Design

In this section, I will discuss research that provides some insight about students' abilities to design statistical studies. In carrying out the present study, special attention was paid to whether or not issues encountered in the research described in this section arose

among the students studied. Since the focus of the present study is upon the high school level, the research literature discussed below includes descriptions of the thinking of students at or near the high school level.

Watson and Moritz (2000a) investigated 3-11 grade Australian students' abilities to detect bias. They found that students ranged in sophistication from those who offered no criticism of situations in which bias would naturally occur to those who recognised the need for samples to be representative and unbiased. The ability to detect bias seems to be related to grade level. When Watson and Moritz (2000a) studied the same students 2 to 4 years later, they found that students tended to improve by one or two levels of sophistication in thinking. The results of the Watson and Moritz (2000a) study show that students do not always recognise that unrepresentative and biased samples produce undesirable results, but that their ability to detect bias seems to improve as they progress through school.

Data supporting the finding that the ability to detect bias is related to grade level occur in U.S. students' response to an item on the 1996 National Assessment of Educational Progress (NAEP). While approximately one-half of eighth grade students responded correctly to an item designed to assess their ability to recognise the potential for sample bias, approximately three-fourths of grade 12 students responded correctly to the same item (Zawojewski & Shaugnessy, 2000). This finding indicates that the ability to detect bias in study design is present more frequently among older students.

In order to design an effective statistical study, it is not sufficient to simply recognise that samples need to be representative and unbiased. One must also use methods with the potential to produce such samples. It is not beyond the ability of students to do so. Watson and Moritz (2000b) conducted a study in which they interviewed Australian students in grades 3, 6, and 9 about their ideas pertaining to sampling. They found that some of the students interviewed understood the roles of randomization and sample size in producing a representative sample. Zawojewski and Shaughnessy (2000) noted that about two-thirds of eighth-grade U.S. students taking the 1996 NAEP could correctly choose the sampling method that would provide the least biased results when given several choices of sampling methods in a multiple choice question. Given these findings, it seems reasonable for students to be expected to develop the ability to choose appropriate methods for sampling during their high school years.

Another essential part of effective statistical study design is deciding when and how to conduct experimental studies rather than non-experimental ones. This can be challenging for college students. Heaton and Mickelson (2002) found that undergraduates had some difficulty matching appropriate data collection methods to the quantifiable questions they had posed for class projects. Derry, Levin, Osana, Jones, and Peterson (2000) described the development of undergraduates' statistical thinking ability in regard to study design. Students in the course studied showed significant gains in knowledge of the design of convincing experiments and the concept of random sampling. Despite the overall gains, many students still tended to confuse the concepts of random sampling and random assignment after the course. Given the difficulties college students have exhibited with deciding when and how to conduct experiments, one would certainly expect experimental design to be a non-trivial matter for high school students.

The research literature described in this section highlights some of the relevant aspects that high school students need to attend to as they design statistical studies. For any given quantifiable question of interest, students must determine whether an experimental or a

non-experimental method is appropriate. Once that determination has been made, formal methods can be used in order to enhance the design of the study. If non-experimental survey methods are appropriate, students need to produce representative samples from the population of interest and eliminate any possible bias. Formal tools such as drawing random samples can be incorporated. If experimental methods are appropriate, experimental designs can be enhanced by the use of formal principles such as random assignments.

Theoretical Perspective

For the present study, the Structure of the Observed Learning Outcome (SOLO) Taxonomy formulated by Biggs and Collis (1982, 1991) was used to help identify cognitive levels in regard to study design. The model has been used effectively by other researchers to help identify various levels of sophistication in statistical thinking. The statistical thinking framework for elementary school students of Jones et al. (2000) and the Middle School Students Statistical Thinking (M3ST) framework (Mooney, 2002) are both based upon the SOLO Taxonomy. In addition, Watson, Moritz, and colleagues have conducted several studies in which Biggs and Collis' cognitive theory was used to describe the relative sophistication of students' responses to statistical thinking tasks (e.g., Watson, Collis, Callingham, & Moritz, 1995; Watson & Moritz, 1999a; Watson & Moritz, 1999b; Watson & Moritz, 2000a; Watson & Moritz, 2000b).

Biggs (1999) described how the SOLO Taxonomy can be used to form a hierarchy of responses to academic tasks. There are five levels in the taxonomy: prestructural, unistructural, multistructural, relational, and extended abstract. Prestructural responses show little evidence of learning relevant to the task at hand. Unistructural responses focus upon one relevant aspect of the task while missing several others. Multistructural responses incorporate more than one relevant aspect, but there is no unifying theme given for the aspects. At the relational level, a unifying theme is apparent along with multiple relevant aspects. Responses at the extended abstract level are "breakthrough" responses that are not just coherent applications of academic learning, but go beyond the task at hand to apply the coherent whole to new areas. The first four levels in SOLO served to help differentiate among levels of response in the present study.

Research has indicated that the unistructural, multistructural, and relational levels cycle and repeat themselves in empirical data (Campbell, Watson, & Collis, 1992; Pegg, 1992). As a consequence, the SOLO taxonomy can be described completely in terms of unistructural-multistructural-relational (UMR) cycles. The middle three levels in SOLO comprise one complete UMR cycle, while the lowest level (prestructural) comprises the relational level in a less sophisticated UMR cycle, and the highest level (extended abstract) comprises the unistructural level in a more sophisticated UMR cycle.

Methodology

A qualitative design was chosen for the study because qualitative designs can be used profitably to investigate intricate thinking processes (Merriam, 1988; Strauss & Corbin, 1990; Bogdan & Biklen, 1992). Within the qualitative design, task-based clinical interviews were used as a primary means of data collection. Goldin (2000) noted that task-based interviews allow researchers to "focus research attention more directly on the

subjects' processes of addressing mathematical tasks, rather than just on the patterns of correct and incorrect answers in the results they produce" (p. 520).

Participants

Purposeful sampling (Patton, 1990) was used in the selection of study participants. The goal of purposeful sampling is to "select information-rich cases whose study will illuminate the questions under study" (Patton, 1990, p. 169). The central questions of this study concerned the identification and description of different levels of statistical thinking. Therefore, a maximum variation sampling strategy (Patton, 1990) was used to select the purposeful sample. A maximum variation sample includes people who have had significantly different experiences in some area, and it allows the researcher to describe patterns of variation within the group studied (Patton, 1990). In this study, students were chosen on the basis of the different types of mathematics courses they had taken while in high school. The goal was then to describe the variation within the responses to interview tasks given by these students. Maximum variation samples are not drawn in order to make statistical generalizations to a larger population, but rather to describe variation and significant patterns within the group (Patton, 1990). Accordingly, in this study, I did not seek to statistically generalise my findings to all high school students, but instead sought to describe the patterns in thinking observed among the diverse sample chosen.

There were three different categories of participants interviewed for the study. The participants in the first category were college freshmen who had recently completed a semester-long high school statistics course. Participants in the second category were college freshmen who had recently completed a year-long high school statistics course. The participants in the third category were still in high school at the time of the clinical interviews. Each of the participants in the third category was enrolled in Algebra, Geometry, Advanced Placement (AP) Calculus, or AP Statistics at the time of the interview sessions. In total, fifteen students participated in clinical interviews. Three of the participants came from the first category, four from the second, and eight from the third.

Students were recruited from each of the three categories by contacting several university instructors and high school teachers. Several instructors at a U.S. Midwestern university helped in the recruitment of students by mentioning my study in their classes and asking for volunteers. The first two categories of students interviewed included students from three of those instructors' classes. The very first category also included a student from a different U.S. Midwestern university who was recruited with the help of her former high school statistics instructor. Teachers at two different U.S. Midwestern high schools helped in the recruitment of students in the last category by discussing my study with their classes and asking for volunteers. Four students in the last category came from one of the high schools, and four came from the other high school. Tables 1 and 2 summarise academic background information about the students interviewed.

Table 1
College freshmen who had recently completed a high school statistics course

Student	Length of statistics course taken in high school
Lisa	One semester
Kristen	One semester
Laura	One semester
Jeff	One year
Hillary	One year
Paul	One year
Julie	One year

Table 2
Students enrolled in high school at the time of the study

Student	Class in high school at time of study	Course enrolled in at time of study
Crystal	Senior	AP Statistics
Bill	Senior	AP Calculus
Luke	Senior	AP Calculus
Jessica	Sophomore	Geometry
Nancy	Sophomore	Geometry
Brooke	Sophomore	Algebra
Rick	Sophomore	Algebra
Daniel	Freshman	Honours Geometry

Note. In general, age breakdowns for high school classes in the U.S. are as follows: freshman = 14-15 yrs., sophomore = 15-16 yrs., junior = 16-17 yrs., senior = 17-18 yrs.

Although college freshmen were included in the study, their patterns of thinking were considered fairly reflective of those one would expect to find among high school students. The college freshmen who participated had all graduated from high school within the past six months. Their interviews were all conducted during the first three months of the academic year. None of them were enrolled in college courses that progressed significantly beyond the statistical content encountered in the AP Statistics course (College Entrance Examination Board, 2001) that is becoming common in U.S. high schools. The college freshmen were included in the study in order to obtain the perspectives of students who had progressed all the way through a statistics course while in high school.

Interview Protocol

The interview tasks that are reported upon in this paper are shown in figures 1 and 2. The tasks represent part of a larger overall interview script (Groth, 2003). The specific content for each of the tasks came from current curricular recommendations for high school statistics courses (NCTM, 1989; 2000; College Entrance Examination Board, 2001; Cobb & Moore, 1997).

Suppose that the governor of Florida puts you in charge of finding answers to the following questions:

(a) What is the typical income of adults in the state?

(b) Will I be re-elected in the election this fall?

(c) What percentage of the state is computer-literate?

(d) Does the new drug for treating the West Nile virus actually work?

(e) How successful was the law which raised the minimum driving age from 16 to 18?

Describe a plan for gathering the information you would need in order to answer each of the questions, and how you would carry out each plan and report the results to the governor

Figure 1. Interview task 1

Suppose that in 1999, a newspaper reporter took a random sample of 15 department stores from the state of Illinois. For each department store he sampled, he found out how much the highest paid man and the highest paid woman in the department store were paid per hour. The reporter wants to repeat this survey again in 2005 and expand it to the population of the entire United States. Describe a plan he could use for getting the information needed for the survey.

Figure 2. Interview task 2

Procedure

Each of the interview participants was informed that it would take a total of approximately 2-3 hours to answer all of the interview questions. Six of the participants in the first two categories decided to split the total interview time between two interview sessions. Julie was the only student in the first two categories who decided to do the entire interview in one session. Most of the participants in the last category were interviewed during study hall periods, so the total interview time for these students was generally split among three different 50-minute study hall periods. Only one of the students in the last

category, Daniel, decided to do the entire interview in one sitting by scheduling an interview session after school hours.

During the clinical interview sessions, an interview protocol (Groth, 2003) was administered in the same order to each of the students. Task 1 (figure 1) was posed at the beginning of each interview session. Task 2 (figure 2) was asked approximately midway through each interview. Task 1d served to elicit thinking about experimental study design, while the remainder of the questions elicited thinking about non-experimental study design. As students were interviewed, I collected data by taking field notes, audio recording responses, and keeping any written work they completed. The audio recordings were later transcribed for analysis.

Data Analysis

Analysis of interview transcripts was advised by the constant comparative method described by Maykut and Morehouse (1994). One of the defining features of the constant comparative method is that data analysis takes place concurrently with data collection. As the data were collected, the responses were examined and grouped according to their relative sophistication. Sophistication was judged in terms of the statistical appropriateness of the response, the number of relevant aspects incorporated in the response, and how well connections were made among the relevant aspects incorporated. After responses had been sorted into groups containing similar patterns of thinking, descriptors were written to capture the essence of each different pattern identified. The end result of the constant comparative method of data analysis was the generation of sets of descriptors for different patterns of thinking regarding both experimental (task 1d) and non-experimental (task 1a, b, c, e and task 2) study design tasks.

Some of the aspects of a check-coding procedure described by Miles and Huberman (1994) were used to finish the process of data analysis. First, a random sample of six students was drawn. Then, two researchers familiar with the Biggs and Collis cognitive framework who had not yet seen the data analysed responses given by each of the six students drawn. The two researchers categorised the student responses according to the descriptors the author formulated during the previously described data analysis process. The two researchers then met with the author to discuss the conclusions they had come to during data analysis. In cases where disagreement occurred about categorizations of students by descriptors or the evidence supporting the formulation of sets of descriptors, the disagreements were discussed until consensus was reached. In cases where disagreements occurred, the originally written descriptors were revised or student responses were re-categorised. Some of the descriptors and categorisations were later further refined on the basis of editorial comments made on an earlier draft of this paper.

Results

The discussion of results is divided into two main parts. The first part describes patterns of thinking observed in response to the tasks lending themselves to non-experimental study design (tasks 1a, b, c, e and task 2). The second part describes patterns of thinking observed in response to the task lending itself to an experimental study design (task 1d).

Designing a Non-Experimental Study

Important statistical thinking skills for high school students include being able to "understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each" (NCTM, 2000, p. 324) and to "know the characteristics of well-designed studies, including the role of randomization in surveys" (NCTM, 2000, p. 324). In tasks 1a, 1b, 1c, and 1e (see figure 1), students were asked to design studies to answer questions of interest about people who live in the state of Florida. No study design was imposed upon them, and they were free to approach the questions in any manner they deemed reasonable. In task 2 (see figure 2), students were to expand a survey that had taken place in one state in order to include the entire United States. Students' responses to the tasks led to the formulation of the descriptors for patterns of thinking summarised in Table 3. In table 3 and the following discussion, students' names are categorised according to the highest level they attained in answering the interview questions pertaining to non-experimental study design (e.g., a student who answered task 1a at a prestructural level and then answered task 1b at a multistructural level would have his/her name associated with the multistructural level).

Table 3
Levels of thinking associated with designing a non-experimental study

Pattern descriptor	Students whose responses reflected the pattern	SOLO level associated with the pattern
Data collection with concern for representativeness and one or more methods to ensure it.	Bill, Crystal, Daniel, Hillary, Kristen, Luke, Paul, Rick	Relational/cycle 2
Data collection with concern for representativeness.	Jeff, Lisa	Multistructural/cycle 2
Data collection without concern for representativeness.	Brooke, Julie, Jessica, Laura	Unistructural/cycle 2
No design strategies articulated, but is aware of the existence of studies and empirical data.	Nancy	Prestructural = Relational/cycle 1

The process of differentiating among levels of sophistication was aided by the SOLO taxonomy and other literature previously summarized in this paper. Responses that discussed the desirability of representative samples were considered more sophisticated than those that did not, since representativeness has been identified by previous research (Watson & Moritz, 2000a; Zawojewski & Shaugnessy, 2000) as an aspect of study design. In addition, responses that incorporated methods to ensure representative samples were considered more sophisticated than those that did not. This distinction was made because of the research illustrating the fact that some students do not have complete understandings

of methods that produce representative samples and when they should be applied (Watson & Moritz, 2000b; Zawojewski & Shaugnessy, 2000).

Prestructural level. The least sophisticated pattern evident in response to designing a non-experimental study was that exhibited by Nancy. She relied primarily on pre-existing studies in order to gather the information needed to answer each question posed. She briefly mentioned interviewing people to determine whether or not the governor would be re-elected in the next election (task 1b), but the idea was not developed. She made use of pre-existing information and studies in books, periodicals, and the internet in order to answer each of the other questions in the tasks. For example, when asked to determine the success of a law that raised the minimum driving age from 16 to 18, she stated,

> Well, it seems like I'm relying on the internet a lot, but that's basically how I would. I guess you'd have to look up the accident claims from insurance companies and see if the claims were higher or lower or whatever after the law was passed.

While she recognised that empirical data would be useful for answering the questions of interest, she relied on others to gather the data for her rather than developing her own data gathering techniques.

Nancy's response to task 2 also lacked reference to any original empirical data gathering techniques. When asked how she would extend the survey described to the entire United States, she stated, "Well, what I would do to make it easier, I would do a study like this [study described in the problem], but I would do one for every state, and then I would probably average them out from there." Her study design incorporated no new aspects beyond those already described in the context of the problem. Further, there was no evidence that she understood the purpose of the data gathering techniques described in the problem context.

Nancy's responses to the study design tasks reflected the prestructural level in the SOLO Taxonomy. Biggs (1999) noted that prestructural responses show little evidence of learning relevant to the task at hand, and are often tautological in nature. Nancy's responses to each part of question 1 gave little evidence that she had developed understanding of how empirical data gathering techniques are relevant to the task of study design. Further, her response given to question 2 was tautological, in that it was essentially a restatement of the problem with no original ideas for study design provided. Hence, Nancy's responses to the tasks involving non-experimental study design were categorized as prestructural.

Unistructural level. In the next pattern identified, students at some point recognised the need for empirical data and also began to develop their own ideas for gathering the data. However, they discussed their data gathering techniques without expressing concern that the data they would gather would be representative of the population from which it was drawn. Brooke, for example, proposed the following plan for predicting whether or not the governor would be re-elected in the election in fall (task 1b):

> In response to the election question, I think I would do some kind of poll that people could answer on the internet or, like, by responding to a phone number in the newspaper. And you also would send door to door scouts out, that ask people if they vote, are they planning on voting for the governor. Then I would probably just display that by a percentage of people that we polled.

Although she discussed a number of data gathering techniques, none of them attempt to ensure that the sample drawn would be representative of the overall population. In fact, the data gathering techniques proposed by students whose responses reflected this pattern were quite likely to produce non-representative samples.

Whitney's approach to determining the typical income of people in the state of Florida (task 1a) further exemplified the use of data gathering techniques without concern for obtaining representative samples. In response to task 1a, she stated,

> I would probably conduct some kind of a survey or a poll just to see, you know, what people say. Or, I don't know, maybe talk to businesses, too. Um, see what they pay their employees, but that would probably be harder. I probably wouldn't talk so much to people as get it written, because then you would have the information to work with if you needed to look back on it, things like that.

Whitney's response to task 1a is similar to Brooke's response to task 1b in that methods of empirical data collection are proposed, but no mention is made of gathering a sample that is representative of the population of interest.

Responses that incorporated data gathering methods without concern for drawing a representative sample from the population were considered to be unistructural. The responses were at least "on track" in the sense that empirical data gathering techniques were proposed at some point. However, the empirical data gathering techniques were the only aspect of relevance included in responses. The important relevant aspect of representativeness was missing.

Multistructural level. Concern for representativeness appeared in responses categorized at the multistructural level. At this level, students proposed data gathering techniques and recognised the importance of obtaining representative samples in sampling situations. Jeff, for example, recognised that it would be important to obtain data for the governor's re-election (task 1.1b) from "a wide range of the population from various backgrounds." In response to the same question, Lisa felt it was important to "make sure that there's all types of people, and that there is no economic bias" within a surveyed sample. However, she did not offer a strategy for obtaining such a representative sample. In general, the students responding within this pattern not only proposed data gathering techniques, but also at some point articulated that it was important for the data gathered to be representative of the larger population.

The study designs proposed in the highest level responses given by both Jeff and Lisa incorporate two aspects relevant to study design: 1) use of data gathering techniques; and 2) concern for representativeness. Since the two relevant aspects were incorporated, but not unified by a coherent strategy for obtaining representativeness, the responses were considered multistructural. The unification of the two relevant aspects came about in responses categorized at the next level.

Relational level. Students giving responses reflecting a more sophisticated pattern proposed methods to ensure that the samples drawn when gathering empirical data would be representative of the population of interest. Bill, for example, suggested the novel strategy of using stratified sampling when asked to expand the department store survey to the entire United States (task 2), saying,

> You could take a direct approach, call each department store, that's just crazy. Or like the census bureau does, you don't have to go to every single house. You could go to, I mean, you could research department store wages in a couple of states, north, east, south, west, all around, and then get an average of that.

Daniel suggested a novel random sampling strategy in response to the same question, saying, "He could pick some cities at random, and pick one man and one woman from a department store from that city chosen at random."

Some students incorporated both random and stratified sampling in the design of studies. Paul's response to the question of determining the typical income of adults in the state of Florida (task 1a) illustrates:

> OK, um, there's a couple of ways, I guess, you could do this. One would be to do a census, you know, and take information from everyone. You'd have to find ways to get to everyone, so they'd have to mail it back, and make sure people - send people out to go find them - that's one way. Another way would be to do a simple random sample, I might stratify the districts, or the counties, I guess. So, you'd have the same amount if there's urban, rural, suburban areas. Then take the same amount - or even it out that way. Then I guess you'd just find the mean salary for each of the adult households.

Responses at this level went beyond the expression of concern for obtaining a representative sample to proposing plausible methods to produce such a sample.

The patterns of response categorized at this final level were relational in nature. They went beyond the multistructural level in that they did not only mention the two relevant aspects of data gathering and representativeness, but also incorporated data gathering methods for producing representative samples.

Designing an Experimental Study

NCTM (2000) recommended that high school students understand experimental studies and be able to conduct them in order to answer quantifiable questions of interest. Researchers have begun to investigate how to design instruction in order to help students learn the principles of experimental design (e.g., Derry, Levin, Osana, Jones, & Peterson, 2000). In order to supplement the current research efforts underway in this area, a task was included on the interview protocol (task 1d) that elicited students' thinking about experimental design. In the task, students were asked to evaluate the effectiveness of a hypothetical drug that had just been developed. Five different patterns were identified in the responses to the task. They are summarised in table 4.

Table 4
Levels of thinking associated with designing an experimental study

Pattern descriptor	Students whose responses reflected the pattern	SOLO level associated with the pattern

Data collection with experimental method and one or more controls to ensure integrity of the experiment.	Crystal, Julie, Lisa, Paul	Relational/cycle 2
Data collection with experimental method.	Bill, Daniel, Kristen, Luke	Multistructural/cycle 2
Data collection without experimental method.	Brooke, Hillary, Jeff, Jessica, Laura, Rick	Unistructural/cycle 2
Relies solely upon pre-existing studies and artefacts.	Nancy	Prestructural = Relational/cycle 1

The process of differentiating among levels of sophistication in response to task 1d was aided by the SOLO taxonomy and by Moore's (1997) discussion of experimental design. Moore (1997) distinguished between an observational study and an experiment, saying "an observational study observes individuals and measures variables of interest but does not attempt to influence the responses" (p. 129), while "an experiment, on the other hand, deliberately imposes some treatment on individuals in order to observe their responses" (p. 129). Moore (1997) went on to say that

> An observational study, even one based on a statistical sample, is a poor way to gauge the effect of an intervention. To see how nature responds to a change, we must actually impose the change. When our goal is understanding cause and effect, experiments are the only source of fully convincing data. (p. 129)

Since task 1d implicitly asked students come to a conclusion about cause and effect, responses that incorporated elements of experimental design were considered to be more sophisticated than those that did not.

Prestructural level. Nancy's response to the task illustrated the least sophisticated pattern of response evident. When asked how she would evaluate the effectiveness of the new drug that had been developed, she stated,

> Well, I guess you'd have to go to like a direct source from…I mean if you couldn't get the answers in a book or a periodical or online or something…If you could actually find someone who had actually encountered the virus and was working on a drug for it, if you had that kind of access.

Hence, she relied solely upon pre-existing studies done by "experts" in order to answer the given question.

Since Nancy's response gives no hint that she would design her own study to collect data, it was considered prestructural. Her response indicated that she would rely solely on studies done by others. No mention was made of what types of methods should be employed for a study to be deemed trustworthy.

Unistructural level. In responses reflecting the next pattern, students proposed their own methods for gathering data in order to answer the question. However, the study designs were observational rather than experimental in nature. Jessica, for example,

proposed gathering information about success and failure from people who had used the drug, stating,

> I would speak to doctors, see what their opinion is. But, probably more importantly, talk to people who have contracted the disease and have taken the drug to see how they felt, have their symptoms got better, and how it worked out for them.

Jessica's response started out very similar to Nancy's in that both initially proposed going to an "expert" source in order to determine the efficacy of the drug. However, Jessica's response was a bit more complex in that she also proposed a data gathering technique that could be used in an original study design. She decided that talking to people who had contracted the disease might yield information pertinent to answering the question at hand.

Brooke's response further illustrated the pattern of thinking in which data gathering methods were mentioned. Like Jessica, Brooke suggested gathering observational data in order to answer the question at hand, stating,

> For the drug one, I think that would be pretty easy, because you would just have to go to the hospital and see how many had been given the vaccine. And then (to see) if it works…you would just have to look at the comments on the chart and see if they had to come back or not. Then you could kind of just make up a little average, I guess, of how many times it did work, and if it was good enough. You would probably have to have a pretty high percentage for it to actually work for most people. So, maybe you would have to compare ages, maybe for like certain ages it works better, or what not.

Responses fitting the unistructural level incorporated data gathering techniques of an observational nature rather than selecting one group to receive the medicine and another not to receive it. The only aspect of design relevant to experimental design incorporated in the responses was the gathering of empirical data. Since this was the only relevant aspect present in the responses, the responses fitting the pattern illustrated above by Jessica and Brooke were considered unistructural.

Multistructural level. The incorporation of a second relevant aspect of experimental design occurred in multistructural responses. Bill suggested testing the drug on animals. Daniel suggested finding some people who had the virus and testing the drug on them. Kristen and Luke both suggested analysing the results of clinical tests without naming specific experimental subjects. Each response reflecting this pattern indicated the recognition of the possibility of imposing a treatment upon a group of subjects in order to answer the question of interest.

The responses indicating recognition that it would be possible to impose a treatment upon a group of subjects in order to determine the efficacy of the drug were considered multistructural. In the responses, the relevant aspect of data collection was incorporated along with the relevant aspect of the imposition of a treatment upon individual subjects. However, the responses simply incorporated the two aspects and did not unify them by suggesting methods by which the integrity of the data obtained by experimental methods could be ensured.

Relational level. Methods for maintaining the integrity of experimental data were mentioned in responses reflecting the relational level. In relational level responses, students not only recognised the possibility of conducting an experiment to answer the question of interest, but they also proposed controls to be put on the experiments in order to help ensure that the experiment would produce trustworthy results. Crystal proposed,

> For this one, I would take a group of people who actually have the West Nile virus, and I would make sure that they are all in the same stage, so that some aren't worse off than others. Then I would

> give part of the people a fake-type drug, and one the real drug, and see how the differences in their improvement turn out.

Julie's strategy for testing the drug further illustrates the relational level pattern. She stated,

> OK, well you would have to do some sort of experiment where you have the actual drug, and then, like a placebo. Two groups are chosen at random and put into random groups. Um, and it would help if it was double blind. And, just have one group taking the new drug and one group taking the placebo.

The pattern illustrated by the responses of Crystal and Julie was considered relational in nature because the two relevant aspects of data collection and imposition of treatment were not just included, but were also tied together by methods aimed at producing trustworthy experimental data.

Discussion

The present study seems to raise at least as many questions as it answers. While it provides a picture of levels of thinking one might expect from students during the high school years, much follow up research remains to be done. It is hoped that the present study will serve as a catalyst for this follow up research. In this final section, I will discuss the limitations of the present study and then conclude by suggesting directions for further research.

Limitations

The picture of students' thinking in regard to experimental design is somewhat limited because of the fact that students answered only one item related to designing an experiment. It is not known how changing the context of the item would have influenced the levels of response exhibited by the students. It possible that an item set in a different context would have prompted students to respond at either higher or lower levels.

The present study does no more than provide snapshots of students' thinking for the purpose of forming a possible hierarchy of levels of response high school students could be expected to give to study design tasks. The amount of data gathered per student is not sufficient to pinpoint the exact developmental level of each student involved. However, the purpose of the study was not to pinpoint the developmental level of individuals, but rather to give a broad picture of different levels of sophistication in response to study design tasks.

It is possible that the levels of sophistication described in the present study could be refined by a similar study involving a larger sample of students. While an effort was made in the present study to obtain data from students representing a range of mathematical backgrounds, it does not claim to have documented all of the characteristics of possible patterns of response that one might encounter.

Directions for further research

Some directions for further research are implied by the limitations mentioned above. For example, one direction for further research would be to describe how changing the context of an experimental design problem changes the level of response elicited from individual students. Another possible direction would be to replicate the present study and include a larger sample of students from various backgrounds. In these ways, the limitations of the present study can serve to spark new research studies.

Another direction for further research would be to investigate the relationship between students' responses to experimental design tasks and their responses to non-experimental design tasks. As mentioned in the limitations, it is beyond the scope of the present study to tie individual students to developmental levels. However, the relationship between the levels at which students' responses are categorised for the two types of tasks in the present study raises some interesting questions. For example, Hillary's name appears at the relational level for non-experimental study design and at the unistructural level for experimental study design. This relationship is turned on its head with Julie, whose name is associated with the unistructural level in regard to non-experimental study design, but at the relational level in regard to experimental study design. Further research might unpack the reasons why some students seem to respond at quite different levels in regard to the two types of tasks.

One more direction for further studies would involve replicating the present study with a younger group of students. It seems that there may be a UMR cycle of lesser sophistication leading up to the ones documented for the high school students. Nancy's responses, classified at the prestructural level, may also reflect the relational level of a UMR cycle in which students gradually become aware of the importance of research but don't develop their own strategies for statistical study design. Interviews with a large sample of younger students could help to determine the characteristics of the levels within the UMR cycle in which Nancy's responses seem to fit.

Acknowledgements

The results related in this paper were part of a doctoral dissertation completed at Illinois State University. I wish to thank my co-chairs Cynthia Langrall and Edward Mooney. I also wish to thank the other two members of my committee, Sharon McCrone and Beverly Hartter.

References

Biggs, J.B. (1999). *Teaching for quality learning at university*. Philadelphia: Open University Press.

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

Biggs, J.B. & Collis, K.F. (1991). Multimodal learning and quality of intelligent behaviour. In H.A.H. Rowe (Ed.), *Intelligence: Reconceptualisation and measurement* (pp. 57-66). Hillsdale, NJ: Erlbaum.

Bogdan, R.C., & Biklen, S.K. (1992). *Qualitative research for education: An introduction to theory and methods*. Boston: Allyn and Bacon.

Campbell, K., Watson, J., & Collis, K. (1992). Volume measurement and intellectual development. *Journal of Structural Learning and Intelligent Systems, 11*, 279-298.

Cobb, G., & Moore, D. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly, 104*, 801-823.

College Entrance Examination Board (2001). *Course description: AP Statistics*. New York: Author.

Derry, S.J., Levin, J.R., Osana, H.P., Jones, M.S., & Peterson, M. (2000). Fostering students' statistical and scientific thinking: Lessons learned from an innovative college course. *American Educational Research Journal, 37*, 747-773.

Even, R. & Tirosh, D. (2002). Teacher knowledge and understanding of students' mathematical learning. In L.D. English (Ed.), *Handbook of international research in mathematics education* (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum Associates.

Fennema, E., & Franke, M.L. (1992). Teachers' knowledge and its impact. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147-164). New York, NY: Macmillan.

Goldin, G.A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A.E. Kelly & R.A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517-546). Mahwah, NJ: Erlbaum.

Groth, R.E. (2003). *Development of a high school statistical thinking framework*. Unpublished doctoral dissertation, Illinois State University.

Heaton, R.M., & Mickelson, W.T. (2002). The learning and teaching of statistical investigation in teaching and teacher education. *Journal of Mathematics Teacher Education, 5*, 35-59.

Jones, G.A., Thornton, C.A., Langrall, C.W., Mooney, E.S., Perry, B., & Putt, I.J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*, 269-307.

Maykut, P., & Morehouse, R. (1994). *Beginning qualitative research: A philosophic and practical guide.* London: The Falmer Press.

Merriam, S.B. (1988). *Case study research in education: A qualitative approach*. San Francisco: Jossey-Bass.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage.

Mooney, E.S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*, 23-64.

Moore, D.S. (1997). *The active practice of statistics*. New York: W.H. Freeman.

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.

Patton, M. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Pegg, J. (1992). Assessing students' understanding at the primary and secondary level in the mathematical sciences. In J. Izard & M. Stephens (Eds), *Reshaping Assessment Practice: Assessment in the Mathematical Sciences Under Challenge* (pp. 368-385). Melbourne: Australian Council of Educational Research.

Strauss, A.L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

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

Watson, J.M., & Moritz, J.B. (1999a). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168.

Watson, J.M., & Moritz, J.B. (1999b). The development of concepts of average. *Focus on Learning Problems in Mathematics, 21*, 15-39.

Watson, J.M., & Moritz, J.B. (2000a). Development of sampling for statistical literacy. *Journal of Mathematical Behaviour, 19*, 109-136.

Watson, J.M., & Moritz, J.B. (2000b). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*, 44-70.

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

Zawojewski, J.S., & Shaugnessy, J.S. (2000). Data and chance. In E.A. Silver & P.A. Kenney (Eds.), *Results from the seventh mathematics assessment of the national assessment of educational progress*. (pp. 235-268). Reston, VA: National Council of Teachers of Mathematics.