National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

AUTHENTIC ASSESSMENT: A LIBRARY OF EXEMPLARS FOR ENHANCING STATISTICS PERFORMANCE

Nancy C. Lavigne
Department of Educational and Counselling Psychology
McGill University
Montréal

© 1994

A thesis submitted to the Faculty of Graduate Studies
and Research in partial fulfillment of the requirements
for the degree of
Masters of Arts
in Educational Psychology

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

THE AUTHOR HAS GRANTED AN IRREVOCABLE NON-EXCLUSIVE LICENCE ALLOWING THE NATIONAL LIBRARY OF CANADA TO REPRODUCE, LOAN, DISTRIBUTE OR SELL COPIES OF HIS/HER THESIS BY ANY MEANS AND IN ANY FORM OR FORMAT, MAKING THIS THESIS AVAILABLE TO INTERESTED PERSONS.

THE AUTHOR RETAINS OWNERSHIP OF THE COPYRIGHT IN HIS/HER THESIS. NEITHER THE THESIS NOR SUBSTANTIAL EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT HIS/HER PERMISSION.

L'AUTEUR A ACCORDE UNE LICENCE IRREVOCABLE ET NON EXCLUSIVE PERMETTANT A LA BIBLIOTHEQUE NATIONALE DU CANADA DE REPRODUIRE, PRETER, DISTRIBUER OU VENDRE DES COPIES DE SA THESE DE QUELQUE MANIERE ET SOUS QUELQUE FORME QUE CE SOIT POUR METTRE DES EXEMPLAIRES DE CETTE THESE A LA DISPOSITION DES PERSONNE INTERESSEES.

L'AUTEUR CONSERVE LA PROPRIETE DU DROIT D'AUTEUR QUI PROTEGE SA THESE. NI LA THESE NI DES EXTRAITS SUBSTANTIELS DE CELLE-CI NE DOIVENT ETRE IMPRIMES OU AUTREMENT REPRODUITS SANS SON AUTORISATION.

ISBN 0-315-99913-6

Canada

Short title:

A Library Of Exemplars

For Enhancing Statistics Performance

ABSTRACT

This manuscript incorporates recent proposals for enhancing the learning of mathematics by developing authentic statistics instruction and assessment for eighth grade students based on a cognitive apprenticeship approach. The goal of instruction was for small groups to create statistics projects that addressed a meaningful research question. To ensure that criteria for assessing such performance were understood, groups were assigned to two treatments--library of exemplars and text--which differed in the degree to which criteria were explicit. The effectiveness of elaborating on criteria through examples (i.e., library) or text (i.e., text) for enhancing learning was examined. Both treatments demonstrated significant performance gains from pretest to posttest. However, students' understanding of representative sampling was significantly better as a result of receiving the library treatment than the text treatment. Making criteria more elaborate through examples of performance can thus enhance students' understanding of more abstract statistical concepts such as sampling.

RÉSUMÉ

Ce manuscrit incorpore des propositions récentes qui visent à augmenter l'apprentissage des mathématiques en développant une stratégie d'instruction et d'évaluation authentique en statistiques pour des élèves de huitième année qui ont vécues une approche d'apprentissage cognitif. L'objectif de l'instruction était de demander à des petits groupes de créer des projets de statistiques qui adressaient une question de recherche significatif. Afin d'assurer la fiabilité des critères d'évaluation selon la comprehension des performances présenter, deux traitements ont été administrer aux groupes qui se différants par l'explicité des critères: librarie d'exemplaires et texte. L'efficacité d'élaborer les critères par exemples (i.e., librarie) ou par texte (i.e., texte) a été examiner selon l'augmention d'apprentissage démontré par les élèves. Les deux traitements ont démontrés des gains de performance significatif sur un examen avant et après instruction. Néanmoins, la compréhension des élèves de l'échantillonage representative étaient significativement meilleur pour ceux qui ont reçue le traitement de librarie que ceux qui ont reçue le traitement de texte. La compréhension des élèves etant exposé à des conceptes abstrait comme l'échantillonage peut donc être augmenter par des exemples qui élaborent les critères d'évaluation.

ACKNOWLEDGEMENTS

My deepest gratitude and admiration to an exceptional scholar and human being, my supervisor, Dr. Lajoie. Thank you for being a constant source of inspiration and for having the foresight to lure me into this research area. Your guidance, continued support, and most of all your belief in me have accelerated my professional and personal development more than you know. To my wonderful parents whose unfailing love, support, and understanding have given me the strength to follow the rainbow: Thank you for believing in my dream. Finally, a special thank you to Steve Munsie whose incredible sense of humour has saved me from taking life too seriously.

I gratefully acknowledge Phil Knox, a gifted mathematics teacher at Lindsay Place High School, for making the data collection possible. A special acknowledgement to André Renaud for his invaluable programming skills. Further acknowledgements to Marlene Desjardins, Cindy Finn, Bryn Holmes, Jody Markow, Steve Munsie, Tina Newman, Litsa Papathanasopoulou, Steve Reisler, and Tara Wilkie for their assistance in the data collection. I would also like to thank Roger Azevedo, Vicki Jacobs, Litsa Papathanasopoulou, and Marguerite Roy for sharing their statistical expertise with me. I also gratefully acknowledge the assistance of Steve Munsie and Roger Azevedo for comments on earlier drafts of the first and last chapters.

The present research was made possible through funding from the Office of Educational Research and Improvement, National Center for Research in Mathematical Sciences Education. A special acknowledgement is due to Thomas Romberg, Director of the Center, for his support of this research.

CONTENTS

Page

Page

Page

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND REVIEW OF THE LITERATURE

Concerns over the quality and characteristics of instruction and assessment have been at the heart of recent proposals for educational reform in North America. Current school practices emphasize discrete and isolated bits of knowledge detached from the context in which such knowledge is used (Hacker & Hathaway, 1991; Resnick, 1987; Shepard, 1991; Whitehead, 1929). Key subtasks are used to assess overall ability despite the restricted and narrow view such measures present of student learning (Frederiksen & Collins, 1989; Hacker & Hathaway, 1991; Moody, 1991). Instructional and assessment practices within this tradition are inconsistent with the recent emphasis of mathematics education on integrated thinking and meaningful learning (National Council of Teachers of Mathematics [NCTM], 1989). Dissatisfaction with traditional classroom instruction and assessment (Shepard, 1989) has led to requests for the integration of authentic classroom activities and assessments as a means of improving the quality of mathematics education (Lajoie, 1991). The term authentic is used to refer to situations that have a genuine relationship to real-world tasks (Moody, 1991). Authentic activities can bring the complexity of real-world problems--problems students are likely to encounter outside the classroom--into the classroom. Such activities provide opportunities for meaningful problem solving. Authentic assessment tasks that reflect the complexity of real-world problems present an interesting challenge to learners. However, whether such tasks succeed in enhancing mathematical thinking and learning has yet to be determined. Research investigating the effectiveness of these activities as well as the validity and reliability of such measures is necessary to determine the potential success of recommended reforms in mathematics classrooms.

Proposals for improving mathematics education reflect the increased value of high-level thinking such as problem solving, reasoning, communication, and connectedness (NCTM, 1989). The objective is to develop instruction and assessment that goes beyond the acquisition of computational and memorization skills. However, this goal can be achieved only by conceiving of mathematics as an ill-structured rather than well-structured discipline. A shift in perspective would greatly reduce the emphasis on memorization skills while increasing the focus on thinking skills (Resnick, 1989). Alternative forms of teaching and assessment that emphasize the active nature of learning can consequently be considered. Reconceptualizing mathematics enables learners to construct their knowledge as well as illustrate and explain their thinking when solving a problem. Such learning is particularly important in the area of statistics where the ability to reason, although crucial for understanding, is rarely engendered (Pollatsek, Lima, & Well, 1981; Tversky & Kahneman, 1971). Providing learners with sufficient conceptual background through instruction at levels earlier than college or university may help redress this problem (Garfield & Ahlgren, 1988; NCTM, 1989; Posten, 1981). However, the abstractness of statistical content can limit younger learners' acquisition of conceptual understanding unless opportunities for active engagement and reasoning are provided.

Student engagement and reasoning can be promoted in project-based environments that situate learners in an authentic context for exploring concepts over long periods of time (Blumenfeld, Soloway, Marx, Krajcik, Guzdial, & Palinscar, 1991). Such exploration can facilitate the acquisition of in-depth understanding. Collaboration and sharing knowledge with peers can also benefit learners in acquiring such understanding (NCTM, 1989; Phelps & Damon, 1989; Wood, Cobb, & Yackel, 1991). However, assessment of learning in group situations poses a challenge. Multiple forms of assessment that examine these skills accurately and provide a complete profile of student and group learning are required. One way to improve the assessment process is to ensure that assessment criteria are valid indicators of student learning and that such criteria are understood by learners. Providing explicit or transparent criteria ensures that students understand performance standards before attempting to attain them (Frederiksen & Collins, 1989). One of the ways in which assessment criteria can be made clear to learners is through a library of exemplars that demonstrates expert performance on relevant assessment criteria (Collins, Hawkins, & Frederiksen, 1991). Such a library can provide clear benchmarks of performance on each criterion thereby ensuring that the criteria are understood. A library of exemplars would be particularly useful for demonstrating abstract statistical content that is rarely understood conceptually. Once internalized, the criteria can be used by learners to assess their own progress in understanding such content (Diez & Moon, 1992).

The purpose of the present study is to implement reform proposals for improving statistics instruction by developing authentic instructional activities and assessment measures to facilitate the learning of descriptive statistics for younger learners (i.e., eighth grade). The effectiveness of these activities and measures is examined in terms of whether (a) statistics instruction facilitates learning at the eighth grade level, (b) projects that integrate instruction and assessment promote learning, (c) small-group cooperation facilitates individuals' understanding of statistics, (d) elaboration of assessment criteria through a library of exemplars enhances performance, and (e) authentic measures reliably assess performance.

Mathematics Reforms in Curriculum and Assessment

Educational standards geared towards improving the quality of mathematics education have been proposed due to students' insufficient conceptual understanding of mathematics. By fostering memorization of computational algorithms and manipulation of symbols, traditional mathematics classrooms fail to promote conceptual or procedural understanding; nor do they emphasize the role of mathematics outside the classroom (Resnick, 1987; Shepard, 1989). As a result, students are not given the power to learn mathematics effectively or to apply their knowledge to real-world contexts. These problems are addressed in reform recommendations for ameliorating mathematical content, instructional conditions, and evaluation of mathematical learning to foster thinking skills such as problem solving, reasoning, communication, and connectedness (NCTM, 1989). These recommendations represent worthwhile or essential mathematical goals which are designed to empower students so that they can acquire confidence and take ownership of their own learning. Classrooms that incorporate real-world problems in problem-solving activities have potential for engaging all students in (NCTM, 1989) (a) formulating and solving problems, (b) making conjectures and developing arguments, (c) validating solutions, and (d) evaluating mathematical claims and evidence. Such activities give learners the power to reason and communicate ideas about mathematical content in problems that are likely to be encountered outside the classroom.

Mathematical Content

Recommendations for improving mathematics instruction emphasize both new and revised content. Proposals for new mathematics content include the commencement of statistics instruction from kindergarten and proceeding through high school (NCTM, 1989). Given that statistical content is highly abstract, activities that make statistical concepts more concrete and provide students with opportunities to apply statistical principles in a meaningful way need to be developed. The purpose of revising content and initiating statistics instruction at the precollege level is to promote student inquiry, investigation, analysis, and interpretation rather than limit learning to the acquisition of computational and algorithmic skills. This goal can be achieved by creating environments that empower students through activities that (a) involve meaningful and realistic problems, (b) allow all students to experiment with and explore statistical concepts more extensively (American Statistical Association [ASA], 1991), and (c) encourage students to problem solve, reason, communicate about statistical ideas, and make connections among concepts. The underlying assumption is that the content to be learned is fundamentally connected with the ways in which it is learned (NCTM, 1989).

Instructional Conditions

Instructional conditions or the ways in which content is learned are determined by the learning environment and the kinds of tasks students are required to do. Proposals for improving instructional conditions in mathematics include developing learning environments that establish mathematical thinking as the classroom norm and encourage students to make sense of mathematics (Lampert, 1990). Mathematical sense-making can be achieved in tasks that require the use of resources and/or tools that reduce the cognitive load (Lajoie, 1993; Salomon, Perkins, & Globerson, 1991) by supporting thinking skills (Lajoie, 1991; NCTM, 1989).

Resources that reduce cognitive load and support thinking skills include peers who work together in groups on problem-solving tasks or on projects that are presented orally (NCTM, 1989). Such collaborative work enables students to share their knowledge which allows them to clarify, elaborate, and justify their ideas. Group work exposes learners to different points of view which provokes thought and challenges beliefs (Collins et al., 1991). In this way, collaboration on problem solving tasks and on oral presentations can support and encourage statistical thinking.

Statistical thinking can also be supported by tools or technologies such as calculators, statistical packages, and graphing utilities. Such tools can reduce memory load by performing lower level operations (Salomon et al., 1991) which allows learners to engage in cognitive activities that would otherwise be out of reach (Lajoie, 1993). For example, computer software that computes statistical analyses provides learners with opportunities to interpret data and to reason about results rather than doing endless calculations which often excludes the possibility of further exploration. Guidance from teachers in using these tools to interpret results is also required. Additional tools that support learning and allow for exploration and communication of statistical concepts include visual and graphic representations such as diagrams, graphs, tables, and concrete models.

Evaluation of Mathematical Learning

The proposed changes in instructional content and conditions also refer to assessment. The relationship between instruction and assessment is a recursive and interdependent one, with learning providing the linking construct. This relationship is important given that learning, while a by-product of instruction, also serves as input for assessment. Assessment in turn, provides information about instruction so that the latter can be adapted to the learning of individual students. While assessment can occur at the end of instruction, the emphasis in the *Curriculum and Evaluation Standards* (NCTM, 1989) is on ongoing or dynamic assessment. Dynamic assessment refers to evaluation that occurs while learners are in the process of solving problems rather than after the problem is completed (Lajoie & Lesgold, 1992). Dynamic assessment thus provides instructors and students with immediate feedback. In this sense, assessment is a guide to learning rather than a terminal point to learning (Diez & Moon, 1992). By representing instructional and learning objectives, assessment can help teachers to make instructional decisions (Collins et al., 1991) and aid learners to assess their own performance (Costa, 1989; Frederiksen & Collins, 1989). It is at this point that instruction and assessment mesh and become integrated. The importance of such integration is discussed in the next section.

Assessment of Mathematical Learning: From Traditional to Authentic Measures

One of the fundamental issues behind educational testing is whether tests represent worthwhile knowledge and mastery. What is deemed worthwhile, however, depends on the type of learning that is valued in education. According to current perspectives, worthwhile knowledge and mastery consists of the ability to problem solve, reason, communicate about ideas, and make connections rather than to memorize (Archibald & Newman, 1988; NCTM, 1989; Romberg, Zarinnia, & Collis, 1990; Wiggins, 1992). Assessments must be designed to promote and measure such skills. Traditional assessments[1], however, were developed based on two sets of purposes which resulted in the promotion of rote learning: (a) classroom purposes which were determined by teachers (i.e., internal assessments) and (b) other purposes which were determined by external sources such as administrators, policy makers, test developers, etc. (i.e., external assessments). Reliance on external assessments that emphasized lower-level skills and rewarded high test scores resulted in the distortion of classroom instruction and assessment (Brandt, 1989;

[1] Traditional assessment in this manuscript refers to all measures characterized by a multiple-choice format which includes standardized and norm-referenced tests.

Frederiksen & Collins, 1989; Hacker & Hathaway, 1991; Kirst, 1991; Shepard, 1989). Instruction became characterized by drill and practice and classroom assessment consequently became grade-oriented rather than learning-oriented (Brandt, 1989; Moody, 1991; Shepard, 1989; Wolf, Bixby, Glenn III, & Gardener, 1991). These practices promote rote learning rather than conceptual understanding. Traditional assessments are thus inadequate for assessing and fostering the type of learning currently valued. Alternative measures that provide rich information about learning through problem solving, reasoning, communication, and connectedness are therefore required (Archibald & Newman, 1988; Lajoie, 1991; NCTM, 1989; Romberg et al., 1990; Wiggins, 1992). Such measures fall under the rubric of "authentic assessment" (Hacker & Hathaway, 1991). The advantages of authentic assessment are highlighted in a comparison of traditional and authentic assessment which is presented in the following sections. The underlying assumptions and testing practices of each type of assessment are discussed in terms of their consequences on learning. This discussion is followed by an examination of design issues relating to the development of authentic measures.

Assumptions and Practices: Traditional vs. Authentic Assessments

In this section, the assumptions (i.e., decomposability and decontextualization) and testing practice (norm-referencing) underlying traditional assessment will be contrasted with the assumptions (i.e., holistic and contextualization) and practice (criterion-referencing) underlying authentic assessment.

Decomposability vs. Holistic Assumptions

Traditional assumption: Decomposability. The decomposability assumption refers to the notion that thought is composed of independent pieces of knowledge and skills (Hacker & Hathaway, 1991). Within this framework, assessments need only examine achievement on key subtasks which are assumed to *reflect* overall ability. Traditional assessment is based on the decomposability assumption. Such an approach to testing, however, results in a restricted and narrow view of student learning (Frederiksen

& Collins, 1989; Moody, 1991). Overall ability cannot be directly assessed through tests that present a snapshot of student learning at one point in time. Individuals have a vast array of knowledge and abilities ranging from low-level skills such as computation to higher-order thinking skills such as problem solving, reasoning, and communication. By excluding complex problems, traditional one-answer or multiple-choice question tests do not provide students with opportunities to develop the latter type of skills (Wiggins, 1990). Rather, the type of knowledge generally assessed by traditional measures is almost strictly declarative. In statistics, for example, knowing what a "sample" is or what it means is assumed to be indicative of how an individual will perform on a task that requires sampling. However, such questions fail to tap into the processes that are involved in such an endeavor, such as planning and reasoning. Rather, restricting measurement to discrete bits of knowledge results in rote learning and in an emphasis on the right answer. By emphasizing correctness above all else, the message traditional tests convey to students is that they are powerless to show what they know. Rather, they must choose or guess someone else's right answer (Hacker & Hathaway, 1991). Students consequently come to believe that explaining one's understanding has no value in learning: The correct solution is what matters. Students are thus portrayed as passive rather than active learners (Kirst, 1991; Shepard, 1989). In essence, instruction based on the decomposability assumption stifles student creativity and insight thereby serving as a barrier to thinking and learning (Hacker & Hathaway, 1991; Romberg et al., 1990).

Authentic assumption: Holistic. The holistic assumption refers to the notion that valid assessments of student progress are obtained when all skills (lower and higher levels) required in performing an activity are measured directly (Hacker & Hathaway, 1991; Moody, 1991). Authentic assessments are therefore holistic in the sense that they provide information about learning that encompasses all aspects of thinking rather than a single skill or subskill. By requiring that learners solve open-ended and complex problems, authentic assessment promotes higher-order thinking skills (Hacker & Hathaway, 1991). Learners are consequently provided with opportunities to explain and justify their solution rather than to select an answer that was generated by a teacher or test developer. By emphasizing quality rather than correctness of a response, authentic assessment conveys to students that their opinions and thoughts are important. Students are thus given an opportunity to articulate, evaluate, and revise their own thinking rather than identify the teachers' knowledge. Learning is thus construed as an active process in which students perform complex tasks rather than engage in recognition or recall on less complex tasks (Linn, Baker, & Dunbar, 1991; Wiggins, 1990). Authentic assessment engages students in sense-making activities (Nitko, 1989) that require the interpretation rather than recognizing the correct answer. Such activities give the control and power over learning back to the students.

Decontextualization vs. Contextualization Assumptions

Traditional assumption: Decontextualization. Decontextualization refers to the assumption that each component of a complex skill is unaffected by the context in which it is used (Hacker & Hathaway, 1991). This assumption has been severely criticized on the grounds that it ignores the integral role played by the physical and social environments in the acquisition and application of knowledge (Greeno, 1989; Resnick, 1987). According to de Lange (1991), phenomena have no meaning except in the context for which the knowledge construction occurs. On this basis it is argued that performance on general achievement measures such as standardized tests, are poor predictors of performance on tasks requiring inquiry, knowledge integration, and communication (Archibald & Newman, 1988). Traditional tests as they currently exist, are not predictive of more authentic forms of achievement.

Authentic assumption: Contextualization. Contextualization refers to the assumption that the context in which learning occurs is a critical component of both the learning and assessment process (Hacker & Hathaway, 1991). This context includes both the physical (i.e., objects in the environment, tools, etc.) and social environments (i.e., interactions with peers and teachers). The contextualization assumption reflects the holistic nature of authentic assessments in the sense that their purpose is to obtain comprehensive information about learning. As such, all aspects of learning and all factors impinging on learning must be examined. The contextualization assumption thus reflects the complexity of real-world learning and enhances the validity of authentic measures.

Norm-Referenced vs. Criterion-Referenced Testing Practices

Traditional practice: Norm-referencing. Traditional measures, particularly standardized tests, present intelligence as fixed, ranked, and predictable (Archibald & Newman, 1988). As such, a single score is used to represent overall ability despite the fact that it does not provide useful information about student learning. Students' scores are ranked and then compared to those of a norm group; i.e., peers. This practice is referred to as norm-referencing (Archibald & Newman, 1988; NCTM, 1993). However, norm-referenced tests do not indicate whether students are doing better or worse (Moody, 1991) but rather treat assessment as a matter of pure measurement (Resnick, 1987). Such tests have been more effective at predicting who will achieve and in describing achievement than helping teachers adapt instruction to enhance the learning of individual students (Linn, 1989). Moreover, by using norm-referenced criteria, traditional tests ensure that at least half the students perform successfully. This practice yields rankings that do not reflect performance in normal learning situations (Wiggins, 1990). In addition, the implication of falling significantly below the norm is that such failure is natural and indicative of an incapacity to learn (Wolf et al., 1991). Learning is therefore perceived as belonging to a select few with the power to learn lying outside of the learners' control.

Authentic practice: Criterion-referencing. The underlying assumption of authentic assessment is that every student has the ability and power to learn. As such, students' scores are not compared to a norm group but rather to preset criteria that reflect agreed-upon learning goals. The generic term for this practice is criterion-referencing. In the case where performance is compared against criteria that reflect NCTM (1989) proposals, the practice is referred to as standards-referencing (NCTM, 1993). By comparing performance to a standard set of criteria that apply to all students, authentic assessment can indicate whether students are doing better or worse (Moody, 1991). Students can compare their performance against the same criteria at various intervals which provides an index of improvement for each student. Authentic assessments use measurement as a tool for enhancing and empowering student learning through improved assessment and instruction.

Summary

The emergence of authentic assessment arose in response to criticisms about assumptions and practices underlying traditional measures. The advantages of more authentic assessment include (a) assessing performance and knowledge directly rather than through other related skills or knowledge, (b) emphasizing higher thinking skills, personal judgment, and collaboration rather than low-level skills, (c) emphasizing the active nature of learning through involvement and participation rather than restricting students to the passive reception of information, and (d) establishing genuine intellectual standards that challenge learners rather than establishing norms that disempower learners.

Issues to be Considered in Authentic Assessment

The advantages of authentic assessment make it an appealing framework for addressing reform proposals designed to improve mathematics education. However, various design issues need to be considered before authentic assessment can be implemented in mathematics classrooms. These include (a) operationalization of authentic assessment, (b) validity, and (c) reliability and cost.

Operationalizing Authentic Assessment

Operationalizing authentic assessment requires a distinction between performance assessment and authentic assessment. The emphasis on performance-based learning has led to confusion regarding the difference between these two types of assessments. The terms performance assessment and authentic assessment are often used interchangeably in the literature yet, they are not synonymous (Meyer, 1992). Performance assessment refers to the examination of a kind of response generated by the student whereas authentic assessment refers to the context in which that response is performed. Authentic assessment is a form of performance assessment, however, not all performance assessments are authentic. It is therefore important to specify in what respects assessment is authentic. Authenticity has many facets, some of which include the following: (a) stimuli, (b) task, (c) complexity, (d) locus of control, (e) motivation, (f) spontaneity, (g) resources, (h) conditions, (i) criteria, (j) standards, and (k) consequences. There is much room for operationalizing authentic assessment given these possibilities. A framework for operationalizing authentic assessment in terms of NCTM (1989) reform proposals has been devised by various researchers (Diez & Moon, 1992; Frederiksen & Collins, 1989; Lajoie, 1991; Linn et al., 1991; Wiggins, 1989, 1992). The following list is a compilation of these frameworks.

1. **Learning.** Authentic assessment must provide multiple indicators of learning. These include both cognitive and conative (i.e. motivation and volition) dimensions of learning.
2. **Tasks.** Authentic assessment tasks must be relevant, meaningful, and realistic. They should also be based on performances and not drills. As such, they should promote thinking rather than the acquisition of bits of information. The set of tasks should be representative for generalizations of overall performance to be made.
3. **Content.** The quality of the content should be high such that fundamental concepts are taught. Content coverage should be comprehensive so that the assessment represents the curriculum in its entirety.
4. **Fairness and Equity.** Authentic assessments must consider ethnic/racial and cultural biases, gender issues, and aptitude biases. Such evaluations should also be equitable over

time. Providing students with opportunities for assessing peers as well as themselves can be useful for teachers in ensuring equity for all students.

5. Classroom. Authentic assessment must be an integral part of the classroom. That is, such measures must reflect the learning goals outlined in the curriculum guidelines.

6. Groups. Authentic assessments must include the development of individuals and groups of individuals in order to assess growth.

7. Consequences. Authentic assessments should be evaluated against curriculum guidelines to determine whether instructional goals were attained.

8. Scoring. As guides to learning, criteria should be descriptive rather than comparative. Language such as "excellent," for example, should be avoided.

9. Benchmarks for success. Examples of various levels of performance should be made available to students prior to assessment to ensure that they know what their evaluation is based on.

10. Transparency. Assessments should be clear enough so that students can assess themselves and others with almost the same reliability as the evaluators.

The above framework is global in that many facets of authenticity are emphasized. As illustrated in the above list, developing authentic assessments requires an equal devotion to the design of authentic instructional activities. These in turn, must be congruent with learning goals. Essentially, the issue is one of validity. Authentic assessments must be designed such that higher-order thinking skills are in fact being measured. Care must be taken to ensure that assessments reflect students' true capacities (Hacker & Hathaway, 1991; Stiggins, 1987).

Validity: Different Perspectives on Student Learning

Given that authentic tasks generally involve complex problems in which various abilities are required for problem solution, the use of multiple forms of assessments is crucial. Samples of student work from a variety of sources are required to obtain different views of student learning. Multiple assessments can provide a more complete and detailed profile of student learning as well as increase validity (Collins et al., 1991; Costa, 1989; Frederiksen & Collins, 1989; Lajoie, 1991; Linn et al., 1991; NCTM, 1993; Shepard, 1989, 1991; Wiggins, 1990, 1992). Three forms of assessments provide different perspectives on learning: (a) paper and pencil tests, (b) video, and (c) computers. Determining which form to use depends on the type of knowledge to be assessed.

Paper and pencil tests in the form of multiple-choice or short-answer exams have traditionally been employed to assess students' declarative knowledge. Although these tests may be useful for obtaining information in areas such as history, they are inadequate for assessing problem solving in statistics. However, the use of paper and pencil tests can be extended to include journals in which students critique their own work and record their own evaluations of their performance (Collins et al., 1991). This journal could then be used by the teacher and student as a forum for discussing how each perceives the progress the latter is making. In addition, the journal can be used to assess the development of students' reasoning and critical thinking skills. Students' assessments can reflect their growing understanding of various concepts and encourage them to examine their performance critically.

Video assessment holds promise for evaluating oral presentations, paired explanations, and joint problem solving (Collins et al., 1991). Such activities reflect real-world endeavors by providing learners with an opportunity to externalize their knowledge through articulation, to clarify and explain their understanding, to listen to others, and to defend their beliefs on the basis of available evidence. Video assessments can therefore provide information about high-level discourse. Information about the learning of individual students and groups of students can also be obtained. Such information is critical for assessing growth more comprehensively and for obtaining a broader view of learning (Collins et al., 1991; Frederiksen & Collins, 1989; Lajoie, 1991; Linn et al., 1991; Shepard, 1991; Webb; 1993; Wiggins, 1989, 1992).

Computers have value for assessing understanding. Computers can provide information about transitions in learning by tracking processes as students perform activities (Collins et al., 1991; Lajoie, 1993; Lajoie, Lawless, Lavigne, & Munsie, 1993). Computer assessments thus provide ongoing measures of learning and highlight specific aspects of students' understanding. This understanding can be facilitated by designing computerized instruction that provides learners with dynamic feedback (Lajoie & Lesgold, 1992). More specifically, hints can be used to assist students in their learning when an impasse is reached. Capturing extensive information about learning through computer assessments is therefore possible. In short, assessment forms such as computers, paper and pencil forms, and videos provide a much broader view of student learning than can be obtained with multiple-choice exams alone. Information collected from these mediums is rich in detail which can enhance the validity of assessments.

Reliability and Cost

Given that authentic assessment provides open-ended problems and tasks that measure a wide range of abilities, scoring such measures is difficult and costly (Brandt, 1992; Hacker & Hathaway, 1991; Moody, 1991; Wiggins, 1990). Difficulties in scoring authentic assessments has consequences for reliability. According to Wiggins, multiple judges are required to ensure inter-rater reliability (Brandt, 1992). Enough information of performance on similar tasks collected over time is also necessary for adequate measurement. An estimated six to twenty tasks are required to obtain reliable individual estimates of performance (Herman, 1992). There is also substantial variation across different tasks, each designed to measure the same thing (Herman, 1992; Shavelson, Baxter, & Pine, 1992). Moreover, designing and scoring authentic measures is costly. However, the costs are outweighed by the gains in professional development and student learning (Wiggins, 1990). By being complex, integrated, and challenging, authentic tasks are said to mirror and support good instruction. However, only when authentic assessment is implemented on a larger scale can the true costs and gains of such measures be determined. Additional research is needed to carefully examine the benefits and pitfalls of using authentic assessments.

Summary

The emerging research in authentic assessment has revealed some of the challenges facing teachers and researchers. While frameworks for operationalizing authentic assessment have been developed, more research is required to examine their effectiveness. Issues of validity, reliability, and cost are also of concern. Suggestions for ensuring validity and reliability respectively include (a) developing multiple assessments that provide rich and detailed pictures of learning; i.e., paper and pencil tests, video assessments, and computer assessments and (b) using multiple judges to evaluate learning on a minimum of six similar tasks.

Recommended Mathematical Pedagogy for Statistics Instruction

Theoretically, authentic assessment has potential for addressing NCTM (1989) proposals to promote high-level thinking in mathematics classrooms. This framework represents a shift in the conception of mathematics from a highly structured to an ill-structured discipline (Resnick, 1989). Conceiving of mathematics as an ill-structured discipline opens the doors to forms of instruction and assessment that emphasize the active nature of learning on complex problems. As demonstrated previously in this manuscript, authentic assessment provides such an alternative. Instruction that seems most effective for promoting student engagement and facilitating thinking skills in mathematics, and more particularly in statistics, can be developed based on two theoretical frameworks: the constructivist and situated learning theories. The following sections describe the constructivist and situated learning theories and then discuss how these theories can be operationalized in the classroom via the cooperative learning and cognitive apprenticeship models. This is followed by a discussion of instructional strategies that seem to address concerns in statistics education.

Constructivism

Constructivism refers to the active involvement of students in th. learning process which allows them to become constructors of their knowledge. Knowledge construction can also be facilitated through group interactions which enable students to communicate their knowledge (Vykotsky, 1978). Revision of knowledge can similarly be fostered in group discussions where learners request clarification, elaboration, and justification. The social component to knowledge construction is referred to as social constructivism. Instruction that provides learners with opportunities to be active learners can thus foster thinking skills. Constructivist and social constructivist theories are therefore consistent with NCTM (1989) proposals in their emphasis on learners' construction, verification, and revision of mathematical models through individual, small-group, or whole classroom discussions.

Although instructional conditions that engender active involvement in the learning process are important, more is required to promote knowledge construction. Students need repeated opportunities to engage in-depth problem solving, assessment, and revision of ideas over extended periods of time (Blumenfeld et al., 1991). Engaging in such activities is often difficult when concepts are abstract and tasks are complex. Meaningful problems must be developed such that abstract concepts are made more concrete and understandable. Instructional methods that make abstract concepts more meaningful ground instruction in real-world contexts. A theoretical framework that provides a basis for designing such instruction is situated learning.

Situated Learning

Situated learning refers to thinking and learning that is situated in physical and social contexts (Greeno, 1989). According to this perspective, cognition occurs in relation to objects, individuals, and situations rather than only in one's mind. Instruction is said to be "situated" or "contextualized" if classroom objects which "afford" students with learning opportunities are utilized to ground the meaning of abstract symbols through real-world connections (Resnick, 1989). In this approach, individuals interact directly with objects and materials in the learning situation rather than manipulate symbols which are detached from their referents (Greeno, 1989). Providing such materials is one way of making mathematical concepts more concrete and meaningful for learners. The social community of learners also promotes learning by providing the "situations and perspectives" whereby students can learn from others. Such an environment can foster the development of problem solving and reasoning skills in mathematics. However, frameworks that provide guidelines for implementing the constructivist and situated learning theories in the classroom are needed. Two models that operationalize constructivist and situated learning theories are cooperative learning and cognitive apprenticeship.

Cooperative learning Constructivist (Vygotsky, 1978) and situated learning theories (Greeno, 1989) emphasize the importance of social interactions for promoting problem solving skills in the classroom. Environments that promote such interaction are referred to as cooperative learning (Blumenfeld et al., 1991; Cohen, 1994; Duren & Cherrington, 1992). Cooperative learning is broadly defined as learning that arises when "students work together in groups small enough that everyone can participate on a collective task that has been clearly assigned" (Cohen, 1994, p. 3). This definition encompasses peer collaboration, cooperative learning, and group work. Evaluating the effectiveness of cooperative learning environments depends on instructional objectives which define productive learning (Cohen, 1994). According to NCTM (1989), small-group learning is productive when students engage in high-level discourse on problem-solving tasks. Given the complexity and ill-defined nature of problem solving tasks, sharing knowledge with peers can benefit students providing them with a more comprehensive knowledge base with which to make sense of mathematical concepts. In a sense, cooperation is like solving a puzzle: having several pieces to the puzzle rather than a single one brings you that much closer to solving it. By having students construct and communicate their knowledge, group problem solving activities ensure that learners explain, justify, and negotiate mathematical meaning. Cooperation can thus foster conceptual learning and higher-order thinking (Cohen, 1994; Duren & Cherrington, 1992; Lajoie, 1991; Phelps & Damon, 1989).

Despite the relative success of cooperative learning environments, more research is needed to examine patterns of students' activity in group learning situations over extended periods (Resnick, 1989; Webb, 1993) and to determine the conditions under which small groups promote learning (Cohen, 1994; Webb, 1991). Although peer collaboration is more effective for making shifts in perspective, the mere presence of a peer is not sufficient for effective learning; joint decision-making is necessary (Rogoff, 1991). Group work can result in a reliance on others which may reduce personal responsibility and independent thinking (Blumenfeld et al., 1991; Webb, 1993). To ensure that cooperative learning situations are productive, ways of posing questions must be carefully engineered (Resnick, 1989). Questions can be delivered in the guise of prompts to engage students in high-level discourse and to extend learning (Rosenshine & Meister, 1992). According to Webb (1991), such discourse is fostered in mixed-ability groups with a narrower range of ability (e.g. highs with mediums or mediums with lows) and homogeneous medium-ability groups. Such groupings facilitate active participation, question asking, and consequently learning. However, Cohen (1994) and Webb (1991) highlight the importance of training students to work cooperatively. These findings demonstrate that cooperative learning groups have potential for engaging students in higher-level thinking if the complexity of cooperation is considered when implemented in the classroom.

Cognitive apprenticeship. Cognitive apprenticeship refers to the notion that skilled learners can share their knowledge with less skilled learners to accomplish cognitive tasks. In a cognitive apprenticeship environment, conceptual and factual knowledge can be exemplified and situated in the contexts of its use (Collins, Brown, & Newman, 1989). The cognitive apprenticeship model proposes six methods for designing an optimal learning environment: (a) modeling, (b) coaching, (c) fading, (d) articulation,

(e) reflection, and (f) exploration. The first three methods refer to knowledge acquired in a social context where experts (i.e. teachers) share their knowledge with novices (i.e. students) and guide learning by: (a) modeling their expertise to make their tacit knowledge explicit, (b) providing assistance through hints (i.e., dynamic assessment), and (c) gradually fading the assistance until mastery is attained. As learners become proficient, they in turn can model their knowledge, coach, and fade assistance for less skilled learners. It is at this point that teaching meets learning (Blumenfeld et al., 1991). However, optimal learning requires the development of autonomous thought which can be achieved by employing the remaining three strategies. Articulation and reflection methods are designed to help learners gain conscious access to and control of their own learning. Articulation methods include activities in which students become their own critics by learning to summarize, clarify, and question. Reflection methods enable learners to compare their processes with those of an expert. Exploration is a final method in which learners establish autonomy by engaging in expert-like problem solving and in defining or formulating problems to be solved.

The cognitive apprenticeship framework holds promise for addressing NCTM (1989) concerns about mathematics instruction. However, if this framework is to be applied in statistics, decisions about the type of content to be taught need to made. At issue is the content to be taught given the emphasis on problem solving, reasoning, communication, and connectedness; keeping in mind that the content to be taught is fundamentally connected with the ways in which it is learned. Although there is a call for including statistics in the mathematics curriculum, there is little discussion of what the specific content should be or how it should be taught. The next section reviews the statistical content that has traditionally been taught and examines how it can be revised for high school students.

Statistics Education

Given that statistics instruction has generally been limited to the university level, little research or practical experience exists to guide the implementation of NCTM (1989) proposals at the elementary and secondary levels. However, a framework for developing statistics instruction can be developed based on literature examining statistics education at the university level. This section will discuss the statistical content currently taught at this level while identifying concepts that have been found to be problematic for students. This is followed by a discussion examining instructional conditions that can be implemented in statistics classrooms at the secondary level.

Statistical Content

The consensus among educators and researchers is that statistics education, in its current form is inadequate (ASA, 1991; Mosteller, 1988; NCTM, 1989; Posten, 1981; Shaughnessy, 1992). A large proportion of university students fail to understand elementary statistics concepts (Garfield & Ahlgren, 1988) even after taking several courses (Posten, 1981). Three reasons for the inadequacy of statistics education have been identified: (a) insufficient conceptual background given to students; (b) abstract nature of concepts; and (c) reliance on formal methods of instruction. Students generally receive statistics instruction in college or university without having had any prior exposure to such content (Garfield & Ahlgren, 1988; Posten, 1981). Adult and middle school learners consequently rely on intuitions or opinions which may conflict with reasoning required to understand concepts such as sampling (Jacobs, 1993; Schwartz, Goldman, Moore, Zech, Smart, Mayfield-Stewart, Vye, & Barron, 1994; Tversky & Kahneman, 1971) and probability (Kahneman & Tversky, 1973, 1982; Tversky & Kahneman, 1973, 1983). The abstract nature of statistical content makes understanding the subject matter difficult, particularly if concepts are taught using traditional methods. Given that statistics lacks decades of curriculum work necessary to build up teaching materials (Posten, 1981), instruction relies on formal methods that emphasize knowledge transmission and rote learning. This reliance leads to difficulties in reasoning about sampling and probability and in understanding the mean as a conceptual rather than computational act (Pollatsek et al., 1981). Such difficulties make developing statistical content for grades K-12 problematic, particularly for grades 5-8 where students are expected to learn concepts such as measures of central tendency, measures of variation, population, sampling, and anomalies (ASA, 1991). Instruction that can make such statistical content more meaningful and less abstract is discussed in the next section.

Instructional Conditions

The move towards more meaningful statistics instruction is at the heart of teaching for understanding. This goal can be achieved by developing classroom activities that anchor statistical concepts so that they are more concrete and meaningful. Providing students with direct decision-making experiences (Hamm, 1992) that require "doing" statistics in a way that illustrates everyday applications of statistics can facilitate statistical understanding. According to NCTM (1989) and ASA (1991), "doing" statistics refers to the following: (a) systematically collecting, organizing, and describing data; (b) constructing, reading, and interpreting tables, charts, and graphs; (c) making inferences and arguments based on data and evaluating arguments based on data analyses; and (d) developing an appreciation of statistical methods as a powerful tool for making decisions. Instruction that provides opportunities to collect, organize, represent, and summarize their own data (a) ensures active student involvement (Mosteller, 1980), (b) can improve students' conceptual understanding of the mean, mode, and median (Zawojewski, 1988), (c) facilitates understanding by enabling students to do statistics while demonstrating its use in the real world (Fischbein & Gazit, 1984; Kapadia, 1982; Pereira-Mendoza & Swift, 1981; Shaughnessy, 1982, 1992; Singer & Willett, 1990; Tanner, 1985; Varga, 1982), and (d) empowers students by making statistical concepts more meaningful and less ambiguous (Watts, 1991).

Statistics instruction must be designed to provide students with opportunities to apply their knowledge of statistical techniques to real-world and everyday problems as well as to learn concepts formally (Mosteller, 1980). Instruction can be further ameliorated by considering the frequency and quality of use of statistical principles (Nisbett, Krantz, Jepson, & Kunda, 1983). However, the nature of students' prior knowledge and intuitive statistical notions must be understood if statistics instruction is to be fully effective (Jacobs, 1993; Schwartz, Goldman, Moore, Zech, Smart, Mayfield-Stewart, Vye, & Barron, 1994). Instruction that builds on students' prior knowledge and allows learners to confront their misconceptions directly can foster learning as well as interest and motivation in statistics (Fong, Krantz, & Nisbett, 1986; Jacobs, 1993). Collaboration between researchers who examine the role of knowledge and intuitions on learning and educators who teach young learners is therefore crucial for improving statistics instruction (Shaughnessy, 1992).

In short, instruction that ensures students' active involvement and application of knowledge to concrete or real-world problems reflects NCTM recommendations and is consistent with the constructivist, situated learning, and cooperative learning models. Opportunities for engaging students in statistical activities rather than passively receiving instruction of formal properties must therefore be provided. A strategy that emphasizes active learning in an authentic or real-world context that has potential for implementing reform guidelines is project-based instruction.

Project-based instruction. A project-based learning environment places students in realistic and contextualized environments that require active engagement over long periods of time (Blumenfeld et al., 1991). Constructivist and situated learning notions are therefore inherent to project work. Such an environment can also serve as a macrocontext (Cognition and Technology Group at Vanderbilt [CTGV], 1990, 1992) where a set of interrelated problems can be used to provide individual or collaborative in-depth explorations of mathematical concepts and principles. Projects thus provide students with opportunities to apply and share their knowledge in order to solve a problem or a set of problems. In statistics, for example, designing a mini-experiment allows for investigation of phenomena, inquiry of variables of interest, collection of data, organization of data, and analysis of data. Statistical packages and graphing utilities can be used to facilitate this process as well as to provide opportunities for interpretation and reasoning. Sharing knowledge with a small group to develop and present a mini-experiment allows students to develop a language to explain, for example, the relationship between measures of central tendency and a given data set (Zawojewski, 1988). Such communication ensures that students cooperate to make sense of statistical concepts and promotes conceptual understanding.

Social interactions can also provide opportunities for learning through cognitive apprenticeships as long as students have good models of expert performance to emulate (Williams, 1992). Expert performance can be modeled by actual practitioners in the field (e.g., statisticians and researchers), teachers, and more capable peers. Opportunities to directly observe practitioners in action, however, are rarely provided. One way to deal with this problem is to provide examples of expert performance in a "library of exemplars" (Frederiksen & Collins, 1989). This library can include (a) videotapes demonstrating the types of questions posed, activities conducted, and tools used by actual practitioners and (b) databases illustrating different types of information collected by experts. Examples of expert performance in designing and conducting an experiment can serve as models that provide clear benchmarks of performance to learners. A library of such exemplars can thus provide a context of apprenticeship for generating a research question, collecting data, analyzing data, representing data, and interpreting data.

Description of the Present Study

Many of the proposals for improving instruction and assessment in mathematics classrooms are still hypothetical. Educational reform is meaningless unless it is anchored to empirical findings which serve as a driving force for implementing change in classrooms. The current study examines whether reform proposals, once implemented in a mathematics classroom, enhance learning and statistical thinking. Instructional content was developed to teach descriptive statistics at the eighth grade level. Instructional conditions were designed to foster thinking skills in activities that (a) required students' active involvement in learning statistical concepts and procedures, (b) contextualized statistical concepts in real-world (i.e., authentic) problems and in macrocontexts such as a project, (c) fostered small-group interaction, and (d) modeled and provided guidance on the statistical problem solving process. Instruction was integrated with assessment through computer tools which connected statistical procedures emphasized in the assessment with concepts and computer skills taught and fostered in the instruction. The text and library of exemplars tools aided students in designing a project; a statistics mini-experiment. Both tools provided explicit assessment criteria in order to facilitate learning and planning. However, the library of exemplars elaborated on the criteria by providing clear benchmarks of performance. In order to assess students' wide range of abilities, multiple assessments such as paper and pencil forms and videos were used.

Research Questions

Five main questions are posed in this research: (a) can statistics be taught at the eighth grade level, (b) can instruction and assessment of statistics be integrated to promote learning, (c) does small-group collaboration facilitate the learning of individual students, (d) does the extent to which assessment criteria are explicit (i.e., text vs. library of exemplars approach) make a significant difference in learning, and (e) can authentic measures reliably assess learning. These questions have guided the development of instructional and assessment materials which are described in the next chapter.

CHAPTER 2: METHODOLOGY

Subjects

The sample for this study was largely middle class and ethnically diverse. Subjects were drawn from an eighth grade mathematics class in an anglophone school in the area of Montréal, Québec. Twenty-one students (nine females and twelve males) were divided into 8 groups, each consisting of two to three students of mixed ability in mathematics. Ability groupings were formed randomly by the experimenter based on the teacher's rating (i.e., high=80-100, medium=60-79, low=59-0) of students' average performance on classroom assessments from the beginning of the year.

Design

Groups were randomly assigned to two treatments: (a) a text treatment in which textual descriptions of the criteria for assessing group projects were presented on the computer and (b) a library of exemplars treatment in which the textual descriptions were supplemented with digitized video clips that demonstrated two levels of performance, average and above average, on the criteria in question. Various measures were used for assessing the performance of individual students and groups of students. Individual students were assessed on (a) a pretest and posttest using a split-plot design and (b) two homework assignments using a between-groups design. Groups of students were assessed on (a) a group project in which three types of ratings--experimenters, groups evaluating themselves, and groups assessing other groups--were used in a split-plot design to examine group performance and (b) a structured journal using a between-groups design to examine planning. Additional feedback regarding students' perceptions of the effectiveness of the instruction was provided by a course evaluation.

Procedure

Groups of students participating in the study worked together for fifty minutes for ten days. Each group worked at an Apple® Macintosh™ workstation which was set up in their regular mathematics classroom. On the first day, students were administered a pretest. This was followed by a brief introductory lesson on statistics in which data and graphs from local newspapers were used to elicit discussions about statistical concepts. On the second day, students were shown a video segment of the David Letterman show in which graphs were utilized to demonstrate the use and misuse of statistics. Students were subsequently put into mixed-ability groups of two or three and then randomly assigned to the text or library of exemplars treatment. On the third day of the study, groups began working on a four-day computer tutorial which included a data collection activity, a 10-15 minute lecture, data analysis activities, and data presentation (i.e., graphics) activities. A homework assignment was given after the second tutorial session. Once the tutorial was completed, students were shown the text or library of exemplars stack to describe the goals and criteria for assessing a group project. A second homework assignment was then distributed. Groups worked on their project for three days and subsequently presented their results to the class using an overhead projector and a Liquid Crystal Display (LCD). Presentations were assessed by experimenters and groups who assessed themselves as well as other groups. The length of each presentation varied from 10 to 15 minutes. Throughout the study, six trained experimenters monitored the groups to assist and expose them to multiple points of view. Group learning during the ten-day period was self-documented by groups' entries in structured journals. Students were administered a posttest on the last day of the study. Group interactions and presentations were audio taped and videotaped, however, only group presentations were examined in this study. Similarly, although all computer work was recorded using ScreenRecorder™ (Farallon Computing, 1990), a MediaTracks™ utility, these are not discussed in this study.

Materials

Equipment

A total of eight Macintosh™ computers were brought into the mathematics classroom: a Classic II, a MacPlus, two SEs, a Powerbook 160, a VX, and two IIcis. The types of computers used by students depended on the treatment to which they were assigned since the speed and amount of storage required to run the information presented in each treatment differed. Lower-end models such as the Classic II, MacPlus, and SEs had 2-megabytes of RAM (Random Access Memory) and 2-megabytes of storage capacity. Such speed and storage capacity was sufficient for running software that presented textual information in the text treatment. Conversely, higher-end computers such as the Powerbook 160, VX, and IIcis ranged from 4-to-8-megabytes in speed and 25-megabytes in storage capacity. Such models were necessary for displaying the visual and auditory information provided in the library of exemplars treatment since at least 4-megabytes of RAM was required to run the digitized video clips.

The computers, in addition to desks and supplementary chairs, comprised groups' workstations which were formed along the walls on the left and right sides of the room to ensure that changes to the normal classroom setup were minimal and to separate groups into different treatments. To avoid confounding the experiment, groups in the text treatment were placed on the left side of the room while groups in the library of exemplars treatment were placed on the right side. All workstations were arranged to encourage group interaction.

Instructional Activities

Instructional activities were designed to be congruent with (a) learning goals recommended by NCTM (1989) and (b) statistical content for grades 5-8 recommended by ASA (1991). Table 1 illustrates how some of these recommendations were implemented in this study. The instructional strategies employed in these activities consisted of three components in a cognitive apprenticeship model of instruction (Collins, Brown, & Newman, 1989): (a) modeling; (b) coaching; and (d) fading. These strategies were utilized by the mathematics teacher and six experimenters as well as embedded in most of the instructional activities. Collaboration between the teacher and researchers made it possible for the former to participate as an experimenter in the study and to initiate his

Table 1

Relationship of Learning Goals With Instruction and Assessment

Learning Goals (NCTM, 1989; ASA, 1991)	Statistical Concepts & Procedures	Instructional Activities and Tools	Assessment
Motivational			
1. Become aware of the utility of statistics in the real world	Statistics	Introductory lesson, randomization activity, and library of exemplars tool	Course evaluation
2. Develop an appreciation for statistics	Statistics	Introductory lesson, randomization activity, and library of exemplars tool	Course evaluation
Cognitive			
1. Use of resources and technological tools			
• Promote the use of statistical and graphical software for entering, analyzing, and representing data.	Data analysis and Data presentation	Tutorial and library of exemplars tool	Group project and structured group journal
2. Develop problem solving skills			
• Understand the problem		Introductory lesson and Tutorial	Pretest/posttest, homework assignments, group project, structured group journal
• Identify key factors in a problem	Hypothesis Identification	Introductory lesson and Tutorial	Pretest/posttest, homework assignments, group project & course evaluation
• Formulate questions	Hypothesis Generation	Tutorial and library of exemplars tool	Pretest/posttest, group project, structured journal, and course evaluation
• Gather data	Data Collection: population, sample,	Randomization activity, tutorial,	Pretest/posttest and group project

Table 1 cont'd...

Learning Goals (NCTM, 1989; ASA, 1991)	Statistical Concepts & Procedures	Instructional Activities and Tools	Assessment
	randomization, sample representativeness, and sample size	and library of exemplars tool	
3. Develop reasoning skills			
• Promote organization and representation of data	**Data Presentation:** data and outlier	Introductory lesson, tutorial, and library of exemplars tool	Pretest/Posttest, homework assignment #1, and group project
• Explore and analyze data for source and method of collection for bias	**Data analysis:** data, outlier, mean, median, mode, and range	Introductory lesson, randomization activity, tutorial, and library of exemplars tool	Pretest/Posttest, group project, structured group journal
• Describing Data	**Data analysis:** data, outlier, mean, median, mode, and range	Introductory lesson, tutorial, and library of exemplars tool	Pretest/posttest, homework assignment #1, group project, structured group journal
• Interpret data		Introductory lesson, tutorial, and library of exemplars tool	Pretest/posttest, homework assignment #1, group project,
4. Develop communication skills:			
• by discussing ideas with peers		Introductory lesson, tutorial, and library of exemplars tool	Structured group journal and group project
• by doing oral presentations		Library of exemplars tool	Group project structured group journal
5. Make connections between:			
• statistics and other subject domains		Introductory lesson and tutorial	-
• statistical concepts and procedures		Introductory lesson, tutorial, and library of exemplars tool	Pretest/posttest, homework assignment#2, group project

own lesson within a cognitive apprenticeship framework. Other experimenters, graduate students with intermediate statistical and computer skills, were given the same instruction on how to model, coach, and fade assistance. This section provides a detailed description of how modeling, coaching, and fading was applied in class discussions, a demonstration, and a computer tutorial. Each activity was designed to situate learning in worthwhile tasks that engaged students in problem solving and cultivated their existing knowledge of statistics.

Modeling

The procedures and reasoning used in descriptive statistics were modeled to students in three ways: (a) by the teacher in an introductory lesson where statistical concepts, procedures, and graphs were discussed; (b) by the researcher in an activity which demonstrated *randomization;* and (c) in a four-day computer tutorial in activities that demonstrated data collection, data analysis, data interpretation, and data presentation.

Introductory lesson on statistical concepts, procedures, and graphs.

The mathematics teacher modeled statistical procedures through graphic representations and used this context to introduce statistical concepts. Graphic representations taken from local newspapers and a video segment of a David Letterman show were used to illustrate the following: (a) the importance of statistics outside the classroom in domains such as economics, politics, financial markets, and sports; (b) the multiple uses of statistics such as describing information, making predictions, and promoting commercial products; and (c) the use of graphs for representing data and statistics visually. The misuse of statistics was demonstrated through the comical nature of David Letterman's graphs which reinforced the previous lesson (i.e., newspaper clippings) and prepared students for the tutorial.

Randomization activity.

The experimenter modeled the *randomization* process by randomly assigning groups to computer workstations. One member of each group was asked to pick a number out a hat; odd numbers corresponding to workstations on the left side of the room (i.e., text treatment) and even numbers corresponding to workstations on the right side of the room (i.e., library of exemplars treatment). Since the latter workstations were equipped with colored monitor screens, students tended to converge on these and shun workstations with black and white screens located on the opposite side of the room. This demonstration provided a framework for explaining the purpose and usefulness of randomization.

Tutorial: problem solving activities. The tutorial provided students with opportunities to learn statistical concepts in problem solving activities that (a) modeled the procedures of data collection, data analysis, data interpretation, and data presentation; (b) fostered the application of newly acquired knowledge; and (c) required the use of data analysis and graphics software to support cognitive load. The concepts taught in the tutorial were primarily in the area of descriptive statistics (see Figure 1 for content). The tutorial activities reduced the abstractness of the concepts by addressing them in a way that was meaningful to students. Meaningful problems were created based on their saliency for the cohort of students participating in this study. Such problems included (a) a mini-experiment where students participated in the collection, analysis, and interpretation of data and (b) three database activities that demonstrated data analysis, data interpretation, and data presentation in three different contexts: school grades, weather forecasts, and world income.

A mini-experiment involving the collection, analysis, and interpretation of pulse rate data (adapted from the ASA Guidelines for K-12, 1991) provided students with a set of interrelated problems which were meant to be fun, meaningful, and challenging. Interrelated sets of problems, often referred to as macrocontexts (CTGV, 1990, 1992), were designed to reduce the abstractness of statistical concepts by situating instruction in a context that extended learning. Three tasks were developed to anchor the instruction provided by the mini-experiment activity. The first task modeled data collection in the gathering of students' pulse rates before (i.e., at-rest) and after physical activity (i.e., runners).

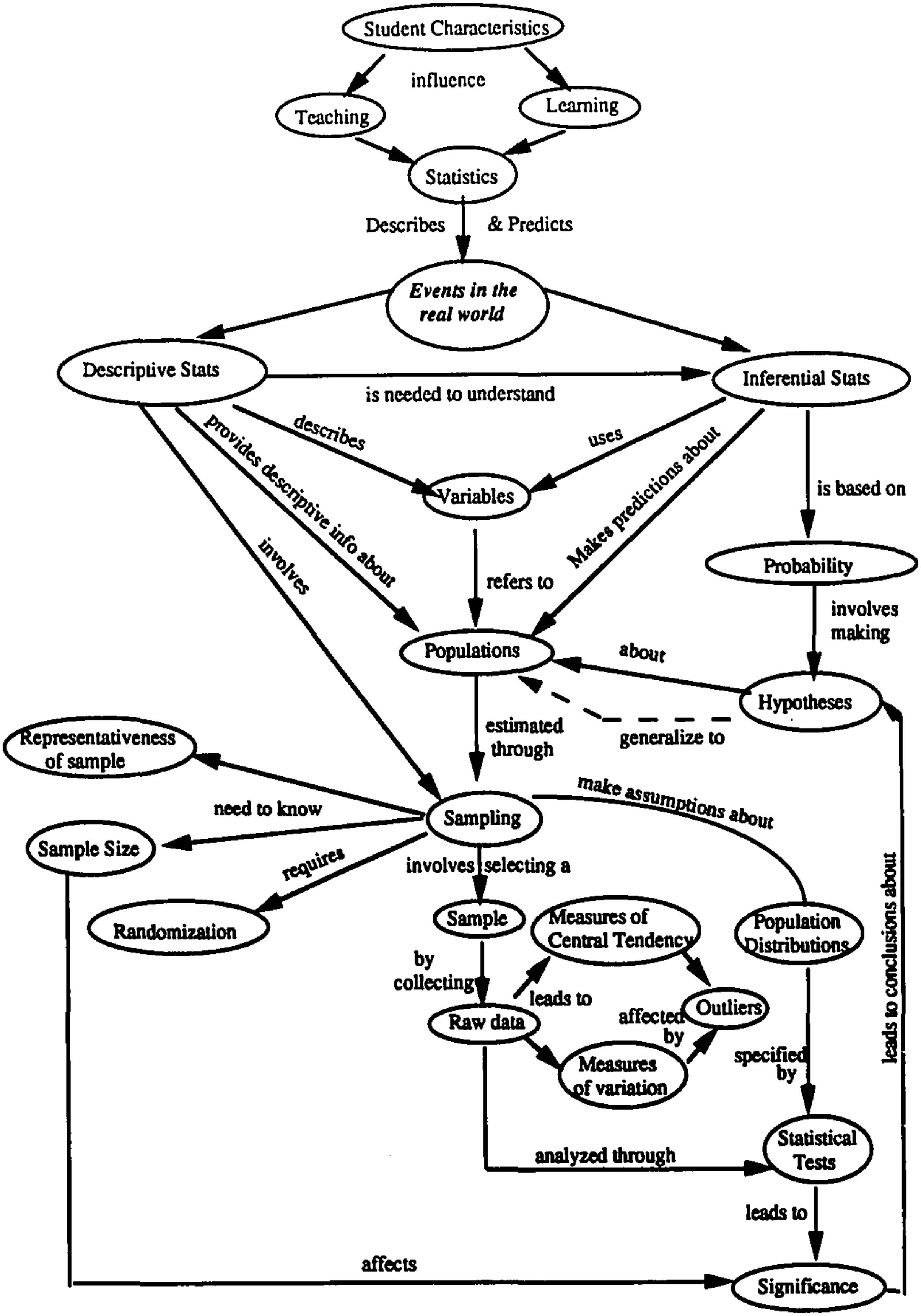
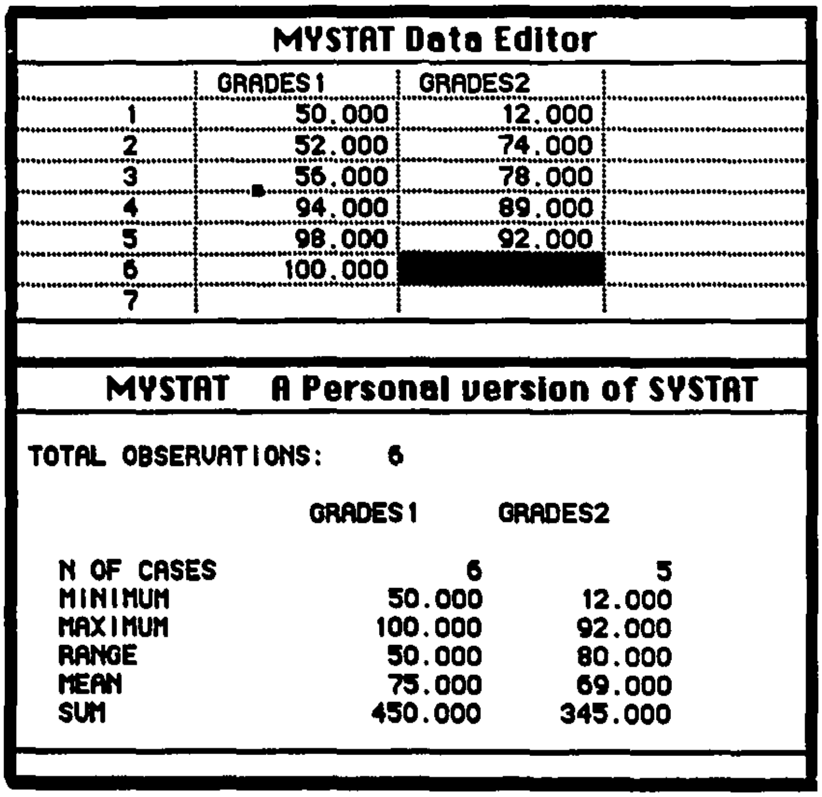


Figure 1. Statistical content.

Once students in each group had collected the "at-rest" data, the tools and procedures of data analysis were modeled using Mystat™ (Systat, 1988). Mystat was used to enter and analyze the "at-rest" data in terms of the *mean* and *range*. Students then collected the "runners" data which was analyzed and compared to the "at-rest data." By comparing the two group means, students could make predictions, test hypotheses, interpret, and revise hypotheses. A 10-15 minute interactive lecture was given within this context to elaborate on concepts more formally and to introduce new concepts. This ensured that students acquired declarative as well as procedural knowledge (Mosteller, 1980) and provided students with opportunities to communicate their knowledge, ideas, and concerns to the teacher, classmates, and experimenters.

The second task in the mini-experiment modeled data analysis by introducing the concepts *mode* and *median* in the context of the "at-rest" pulse data collected by each group. This allowed for a demonstration of the selection of appropriate measures of analysis. Groups were given an opportunity to examine differences between three measures of central tendency and to determine which, according to their particular data, was a better measure and why. The third task modeled data interpretation by comparing "at-rest" data of students in the entire class with each group's data. This activity (a) provided students with an opportunity to apply their knowledge of the mean and range, (b) extended the discussion to *sampling, sample size, population,* and (c) provided another context for discussing randomization.

In addition to the mini-experiment, the tutorial provided three activities based on topics that seemed salient to eighth grade students: school grades, weather forecast, and world income. These database activities (a) demonstrated the analysis, interpretation, and presentation of data and (b) provided new contexts for extending learning and acquiring knowledge. The school grade activity was thought to motivate student interest given the time of year (i.e., end of term). Students entered data specified by the tutorial on the computer. Data analysis techniques were demonstrated through the use of computer software (i.e., Mystat) for entering data and calculating the mean (see Figure 2 for data and analyses). Data interpretation was illustrated through the examination of data presented in charts prior to and following analysis. By requiring that students predict the value of the mean and examine the influence of extreme scores, the school grade activity extended students' learning of the concepts of data and mean and provided a context for introducing the concepts of outlier and sample representativeness.



MYSTAT Data Editor

	GRADES1	GRADES2	
1	50.000	12.000	
2	52.000	74.000	
3	56.000	78.000	
4	94.000	89.000	
5	98.000	92.000	
6	100.000		
7			

MYSTAT A Personal version of SYSTAT

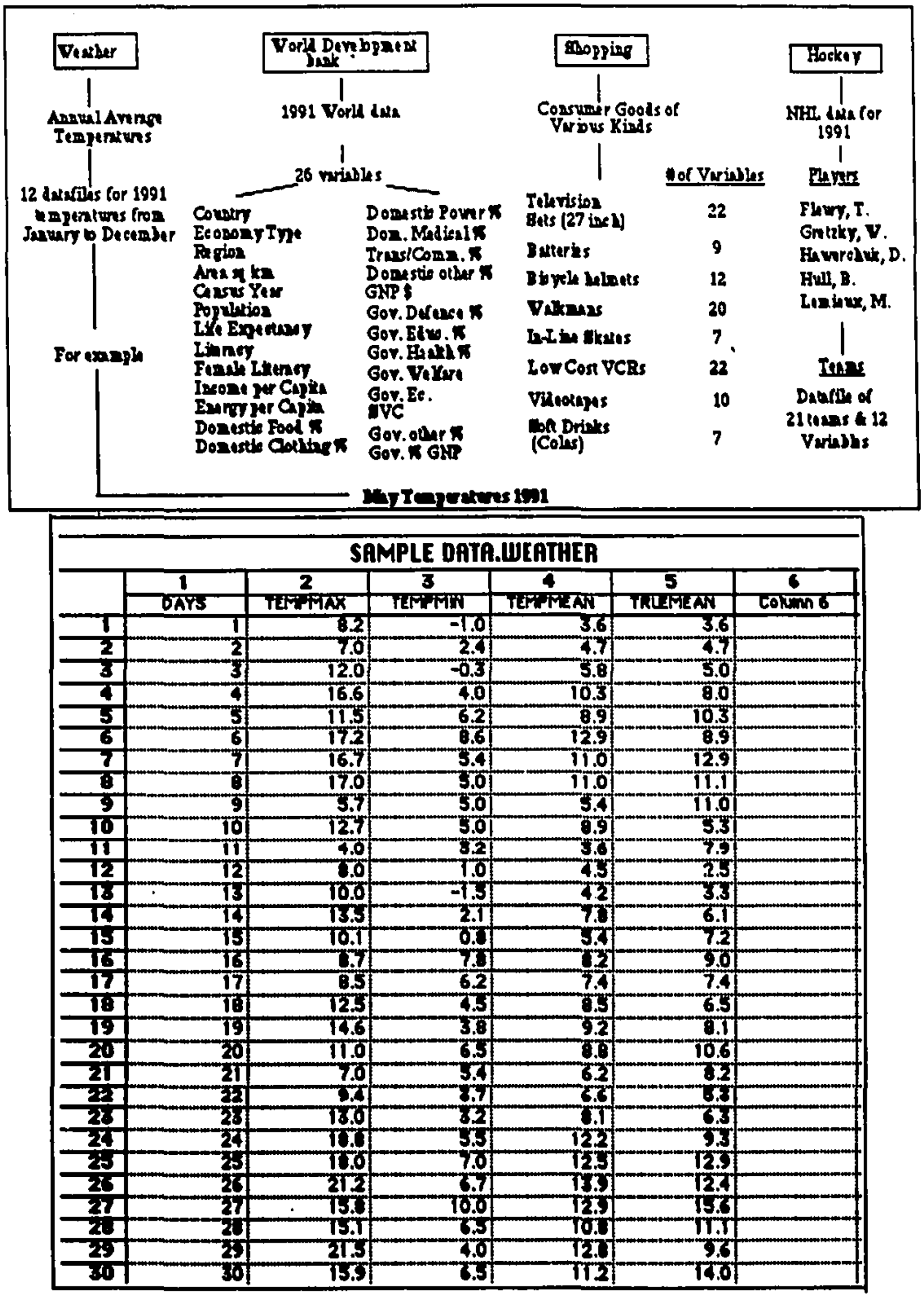
```
TOTAL OBSERVATIONS:     6

                        GRADES1     GRADES2

  N OF CASES                  6           5
  MINIMUM                50.000      12.000
  MAXIMUM               100.000      92.000
  RANGE                  50.000      80.000
  MEAN                   75.000      69.000
  SUM                   450.000     345.000
```

Figure 2. Data and analysis used for the school grades activity in the tutorial.

The weather forecast and world income database activities were thought to motivate student interest since they reflected real-world concerns. These activities provided students with anomalous data and a large pool of variables which allowed for further exploration of concepts and procedures (see Figure 3 for a specification of the variables).



SAMPLE DATA.WEATHER

	1	2	3	4	5	6
	DAYS	TEMPMAX	TEMPMIN	TEMPMEAN	TRUEMEAN	Column 6
1	1	8.2	-1.0	3.6	3.6	
2	2	7.0	2.4	4.7	4.7	
3	3	12.0	-0.3	5.8	5.0	
4	4	16.6	4.0	10.3	8.0	
5	5	11.5	6.2	8.9	10.3	
6	6	17.2	8.6	12.9	8.9	
7	7	16.7	5.4	11.0	12.9	
8	8	17.0	5.0	11.0	11.1	
9	9	5.7	5.0	5.4	11.0	
10	10	12.7	5.0	8.9	5.3	
11	11	4.0	3.2	3.6	7.9	
12	12	8.0	1.0	4.5	2.5	
13	13	10.0	-1.5	4.2	3.3	
14	14	13.5	2.1	7.8	6.1	
15	15	10.1	0.8	5.4	7.2	
16	16	8.7	7.8	8.2	9.0	
17	17	8.5	6.2	7.4	7.4	
18	18	12.5	4.5	8.5	6.5	
19	19	14.6	3.8	9.2	8.1	
20	20	11.0	6.5	8.8	10.6	
21	21	7.0	5.4	6.2	8.2	
22	22	9.4	3.7	6.6	8.3	
23	23	13.0	3.2	8.1	6.3	
24	24	18.8	5.5	12.2	9.3	
25	25	18.0	7.0	12.5	12.9	
26	26	21.2	6.7	13.9	12.4	
27	27	15.8	10.0	12.9	15.6	
28	28	15.1	6.5	10.8	11.1	
29	29	21.5	4.0	12.8	9.6	
30	30	15.9	6.5	11.2	14.0	

Figure 3. Types of databases.

The analysis, interpretation, and presentation of data in both these activities was modeled through the use of a graphical tool called CricketGraph™ (Cricket Software, 1989). This contrasts with the mini-experiment and school grade activity where data analysis and data interpretation were modeled through the use of a data analysis tool called Mystat. In the former case, graphical representations were emphasized while the latter focused on numerical representations for modeling statistical concepts.

In the weather forecast activity the analysis, interpretation, and presentation of data were demonstrated through the use of scatterplots. The maximum, mean, and minimum temperatures of May 1991 were represented by a line which connected the datapoints for each temperature type. These three lines demonstrated the results and provided a graphical representation for interpreting the data. The weather data also provided a new context for exploring extreme scores and the influence of outliers on the mean. A discussion of sample representativeness ensued. Answering questions such as "is the data for May 1991 representative of May weather every year?" could entail further data collection leading to an in-depth exploration of sample representativeness, outlier, mean, and range.

The world income data was used to model the use of pie charts for examining, interpreting, and presenting the relationship between economy type (low, medium, high) and population per country. The data was analyzed and interpreted based on the graphical representation and numerical values (i.e., percentages) which were attached to the levels of a variable. Other variables within this database could be explored by students at their leisure. Two additional databases, National Hockey League (NHL) statistics and consumer reports (see Figure 3 for further information on types of variables for each database) which were designed to motivate individual interest were also made available for students to explore at their leisure or for developing group projects.

Coaching

Introductory lesson on statistical concepts, procedures, and graphs.

Coaching was provided by the mathematics teacher during his lesson on statistics and graphic representations using newspapers. During this lesson, students were encouraged to identify and interpret graphs presented in newspaper clippings based on knowledge they had acquired in other courses. The teacher prompted students with questions and lead-on sentences that cued them for answers. In addition, the teacher tried to closely connect his lesson with concepts mentioned in the pretest and with what students had learned in his history class and in other courses, such as geography and science.

Randomization activity. The randomization activity provided an ideal context for questioning students about the consequences of not randomly assigning groups to workstations since students did not want workstations with the black and white computer screens. Students were asked to explain why they thought such a procedure was necessary and were coached as responses were made. Students were also prompted to identify consequences of not performing randomization and to provide examples.

Tutorial: problem solving activities. Coaching was provided by the mathematics teacher, experimenters, and by prompts which were embedded in the tutorial. The teacher and experimenters facilitated students' learning by coaching them while they worked on instructional activities. By floating from group to group, the teacher and experimenters provided students with multiple perspectives. Students were able to draw on the expertise of each of these mentors.

Prompts which were embedded in the tutorial provided students with a form of coaching (see Appendix A for summary of the tutorial activities and the prompts associated with each activity). Prompts were delivered as questions to (a) ensure that students engaged in productive learning while working in groups (Resnick, 1989), (b) encourage students to reason about data and graphs, (c) facilitate discussion of statistical concepts, and (d) extend students' learning beyond the information given (Rosenshine & Meister, 1992). Students were also coached on how to perform analyses using computer software as well as on how to create and modify graphs for clarity and coherence.

Fading

Tutorial: problem solving activities. Given that much of the learning took place during the tutorial and that this activity was of a long duration (4 days as opposed to 5-10 minutes in the introductory lesson and randomization activity), the strategy of fading was mainly employed during the tutorial. Experimenters, as the "masters" in this learning environment, guided learning throughout the tutorial and gradually faded their assistance as student "apprentices" mastered computer skills, attained statistical competency, and took responsibility for their own learning. At this point, group members were able to guide their own learning and learn from each other, particularly when the more skilled students within each group took the role of master and guided the less skilled members.

Summary

A cognitive apprenticeship model of instruction was adopted to teach students descriptive statistics through modeling, coaching, and fading on three types of instructional activities: an introductory lesson on statistics, a randomization activity, and a computer tutorial. These activities constituted the knowledge acquisition phase of the study which was designed to provide students with knowledge of facts and :ools required to do their own group projects in the performance phase of the study. Projects provided students with opportunities to (a) do statistics by selecting a problem which required the collection, analysis, interpretation, and representation of data; (b) articulate, elaborate, and clarify their understanding of concepts and procedures; and (c) extend learning through the integration of instruction with assessment. The next section describes how these components were implemented in the study.

Instructional and Assessment Tools

Once the knowledge acquisition phase of the study was completed, students were required to apply and demonstrate their knowledge by developing and presenting a group project that involved conducting a mini-experiment. This project provided students with an opportunity to further develop autonomous thought through articulation, reflection, and exploration (Collins et al., 1989). Exploration was provided by engaging students in expert-like problem solving through the definition and formulation of their own research question. Groups could use databases provided in the computer to construct a question or collect their own data. The project also engaged students in data analysis and graphic representation. Opportunities for articulation were provided in group discussions and presentations where interpretations were communicated to peers and mentors (i.e., teacher and experimenters). Presenting projects allowed groups to articulate their knowledge to a larger audience whose perspectives might differ from their own. Questions posed by peers allowed for elaboration and clarification of ideas as well as for re-evaluations of one's understanding. Such questions thus engaged students in reflection. However, the following must occur if groups are to perform satisfactorily: (a) the assessment criteria upon which groups are evaluated must be understood (Frederiksen & Collins, 1989) and (b) connections between the assessment task (i.e., project) and the instruction must be made explicit.

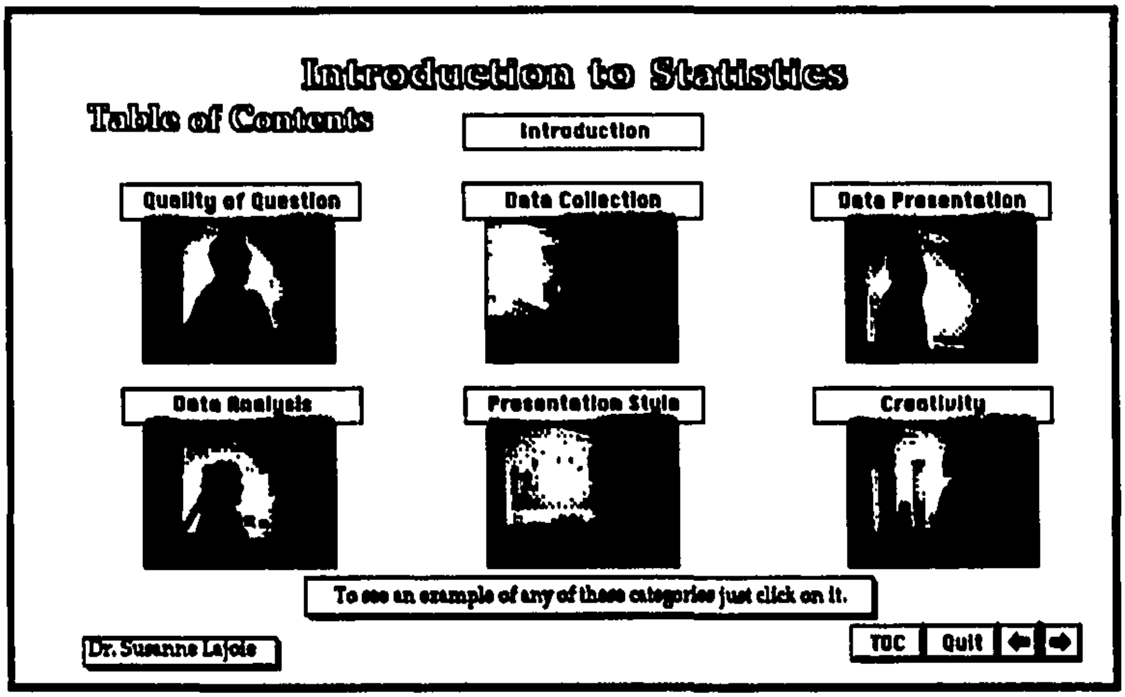
The criteria upon which group projects were evaluated were made transparent by explaining the assessment criteria prior to students' commencement of group projects. In addition, the connection between the instruction and assessment task was made explicit by describing the assessment criteria in terms of the concepts and procedures that were taught in the instruction. More specifically, groups were assessed according to (a) the quality of their research question, (b) the procedures used for data collection (i.e., data, sample, sample size, sample representativeness), (c) the ways in which the data was presented to the class (i.e., graphic representations), (d) type of data analysis and quality of interpretation (i.e., mean, mode, median, range), (e) presentation style, and (f) creativity. Although presentation style and creativity did not refer to any specific concepts, they were useful for assessing the groups' abilities to communicate their knowledge. A 50-point system was developed for assessing the criteria.

Two treatments were developed for administering and making the assessment criteria clear to students: the text and library of exemplars treatments. Computer software, HyperCard™ (Claris Corporation, 1991), was used to administer information about the assessment criteria in each treatment. Details about the software and the degree of transparent information presented in the text and library treatments are provided in subsequent sections.

Text HyperCard™ Stack

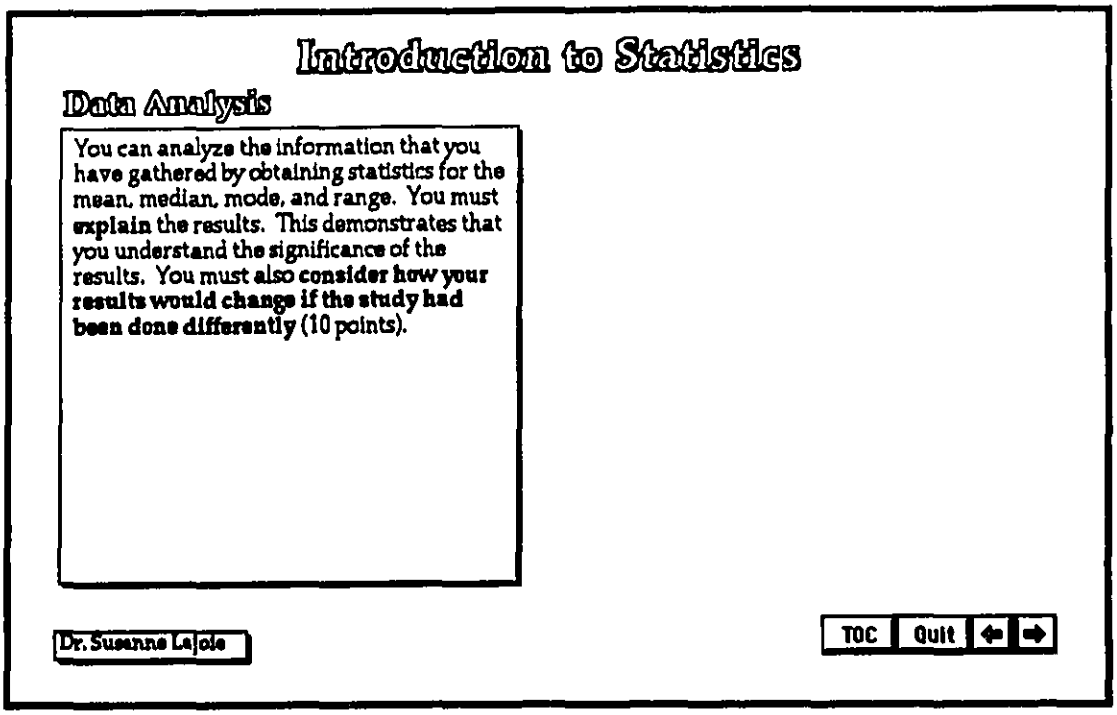
HyperCard is computer software that operates on an Apple® Macintosh™ platform. This software allows users to organize information hierarchically in much the same way information is indexed in libraries. For example, a student inquiring about statistical procedures for conducting experiments would go to the index section in the library and locate the drawer labeled "S " for statistical procedures. All information pertaining to the topic "statistical procedures" is called a "stack" in HyperCard terms. Within this stack are index "cards" which provide specific information about subtopics relevant to statistical procedures. In this study, the "text" stack contained 8 cards which provided specific information about the statistical procedures for conducting an experiment in the authentic statistics project: the table of contents, an introduction to the project, and the six criteria. Textual descriptions of what was meant by each criterion and a specification of the value of each was outlined to groups in every criterion card. Such descriptions provided the explicitness required for understanding the criteria upon which one would be evaluated. The advantage of providing this information on the computer via HyperCard rather than on a piece of paper is that it (a) makes the information more accessible to students, (b) fosters interest in learning the criteria by having students interact with the computer, and (c) allows students to focus on each criterion one at a time and to return to any criterion numerous times.

The text stack was structured such that students could easily access information pertaining to any of the criteria. After reading the introduction regarding the purpose of the stack, students were directed towards the Table of Contents (TOC) where each criterion was specified. To learn about any of the criteria, students merely clicked on the criterion of interest and/or on the arrow keys located on the lower right of the card (see Figure 4). Once selected, information about the criterion is displayed. In the text stack, such information was limited to a description of the criteria (see Figure 5). Students interacted with the information by using the arrow keys and/or selecting the TOC option located next to the arrow keys. This allowed students to move from criterion to criterion at their own pace and to focus on each, one at a time. Students were free to return to the stack at any point during their projects. In this sense, learning was student-controlled. Since the "text stack" made the criteria explicit through textual representations, its effectiveness was compared with that of the library of exemplars in which the criteria were made much more transparent through textual, visual, and auditory representations.



To see an example of any of these categories just click on it.

Figure 4. Table of Contents for the Text stack.



Introduction to Statistics

Data Analysis

You can analyze the information that you have gathered by obtaining statistics for the mean, median, mode, and range. You must **explain** the results. This demonstrates that you understand the significance of the results. You must also **consider how your results would change if the study had been done differently** (10 points).

Figure 5. Example criterion in the Text stack: Data analysis.

Library of Exemplars HyperCard™ Stack

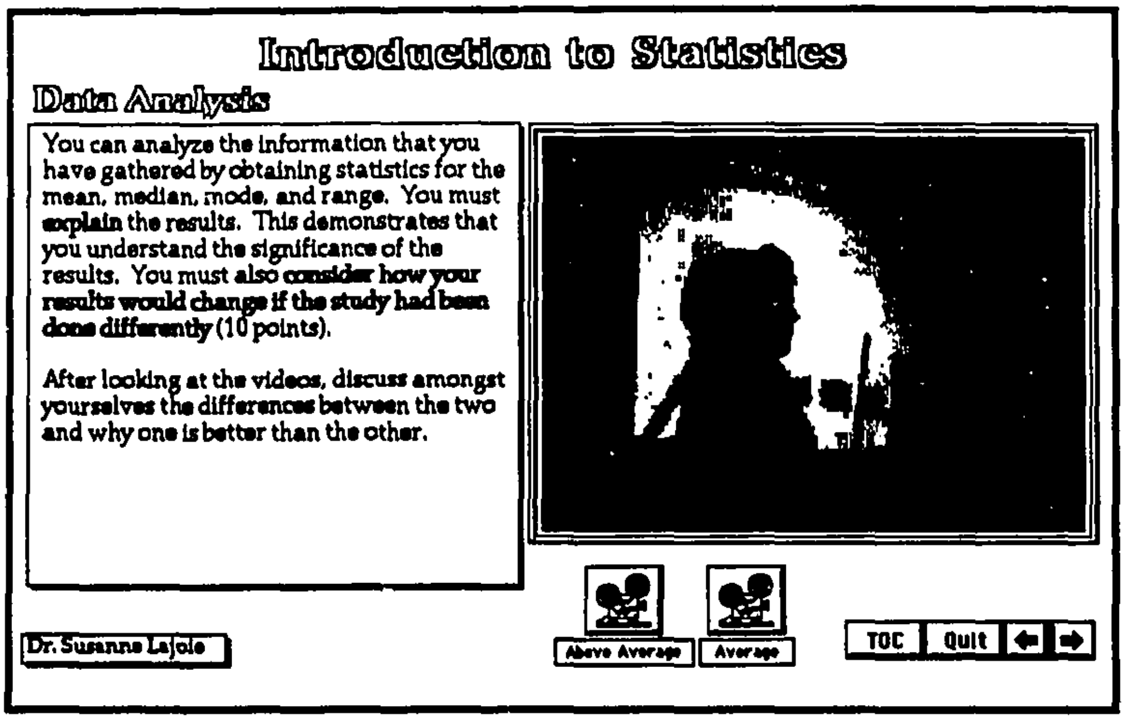
A HyperCard stack similar to the "text " stack was developed to drive student interactions with the library of exemplars. Both the text and library of exemplars stacks represented the assessment criteria textually, however, the library of exemplars went beyond the single text representation that described the criteria to include multiple representations of sound, dynamic movement, and text. Sound and dynamic movement refer to digitized video clips which served to (a) situate the criteria in concrete examples of average and above average performance of students presenting a similar project one year earlier (Lajoie et al., 1993) and (b) provide multiple perspectives of the task by students of equivalent background. In addition, textual representations in the library of exemplars stack went beyond the description of the criteria to include prompts which required students to reason about and discuss differences between the two levels of performances displayed in the digitized video clips (see Appendix B for details regarding the development of the library of exemplars hypercard stack).

Groups using the library of exemplars were first required to read the criteria descriptions, then to look at the digitized video clips, and then they were prompted to reason and discuss the information with students in their group (see Figure 6). Close-up pictures of data presented in the video clips were also included so that groups would have all pertinent information available to them (see Figure 7). The library of exemplars was intended to provide students with benchmarks of performance in addition to the textual descriptions of each criteria as a means of making the learning goals and assessment criteria as clear and as explicit to the learner as possible (Frederiksen & Collins, 1989). The transparency of the criteria should lead to enhanced performance and to more similar group assessments among groups of students and experimenters. The prompts (see Appendix C) were designed to (a) promote students to think beyond the information given (Rosenshine & Meister, 1992), (b) promote reasoning and communication skills, and (c) encourage students to plan their projects accordingly.

Assessment Measures

Assessments in this study were authentic[2] in that they (a) provided multiple indicators to obtain a complete picture of learning and to increase the quality of evaluation, (b) provided students with realistic and meaningful tasks and problems, (c) covered the course content comprehensively by implementing ASA (1991) content recommendations, (d) were equitable by providing students with opportunities to assess peers as well as themselves, (e) were an integral part of the classroom by reflecting learning goals specified by NCTM (1989) and by being congruent with instructional activities that demonstrated relevant statistical concepts and procedures (see Table 1), (f) included the development of both individual and group measures to assess growth and to ensure that group

[2] The measures were deemed authentic based on the compiled framework for operationalizing authentic assessment (see chapter 1).



Introduction to Statistics

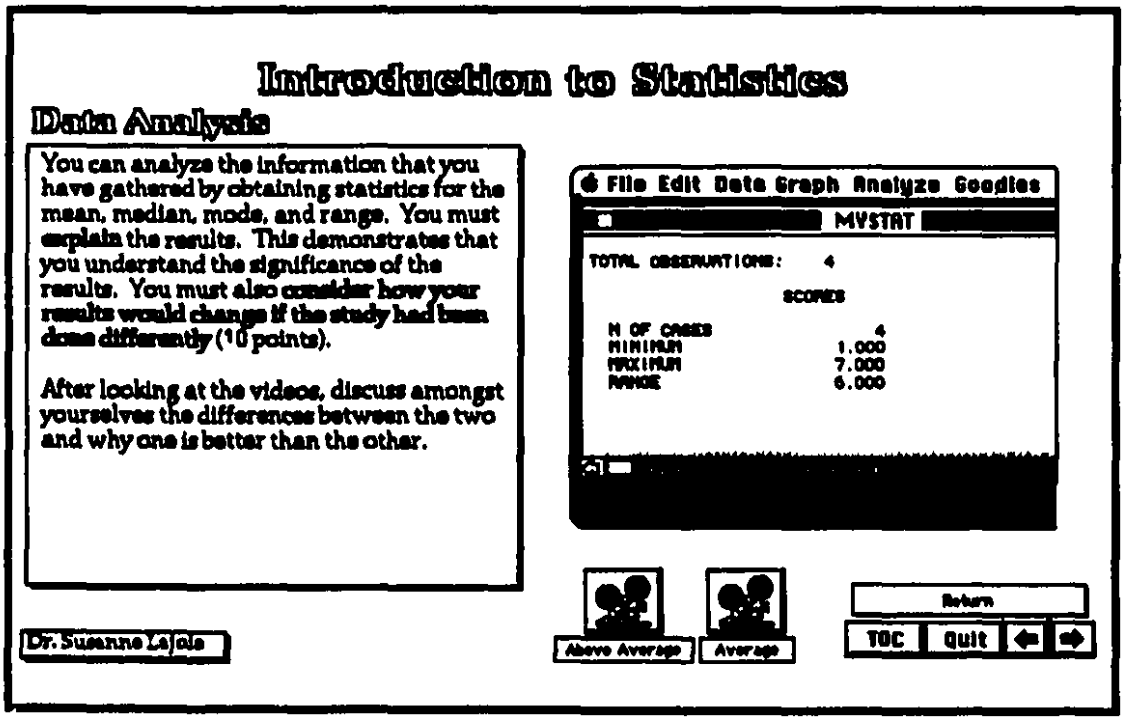
Data Analysis

You can analyze the information that you have gathered by obtaining statistics for the mean, median, mode, and range. You must explain the results. This demonstrates that you understand the significance of the results. You must also consider how your results would change if the study had been done differently (10 points).

After looking at the videos, discuss amongst yourselves the differences between the two and why one is better than the other.

Dr. Susanne Lajoie

Figure 6. Example criterion in the Library of Exemplars stack: Data analysis.



Introduction to Statistics

Data Analysis

You can analyze the information that you have gathered by obtaining statistics for the mean, median, mode, and range. You must explain the results. This demonstrates that you understand the significance of the results. You must also consider how your results would change if the study had been done differently (10 points).

After looking at the videos, discuss amongst yourselves the differences between the two and why one is better than the other.

```
TOTAL OBSERVATIONS:     4

                    SCORES

  N OF CASES                 4
  MINIMUM                1.000
  MAXIMUM                7.000
  RANGE                  6.000
```

Dr. Susanne Lajoie

Figure 7. Example of close-up picture in the Library of Exemplars stack: Data analysis.

assessments do not overestimate individual performance (Webb, 1993), (g) provided benchmarks of performance, and (h) were designed to be transparent to the learner.

The measures provided students with open-ended problems that emphasized the quality of performance rather than the accuracy of solution. This section describes six measures which were designed to provide valid and reliable assessments of performance on (Brandt, 1992; Herman, 1992): a pretest and posttest, homework assignments, oral presentations of projects, assessments of group projects, and a structured group journal. In addition, a questionnaire in the form of a course evaluation was used to provide the experimenter with feedback about students' perceptions of their learning experience.

Pretest /Posttest

The pretest and posttest were designed to measure students' knowledge of statistical concepts and procedures prior to and after instruction. These were completed individually on the first and last day of the study. Since students knew very little about statistics prior to instruction, the posttest was only slightly different from the pretest. Seven of the ten items on the pretest and posttest were the same, however, three pretest items were modified in the posttest (see Appendix D). Numerical values were replaced on one item while the scenario was changed on a second item. In both cases, questions in the pretest and posttest were identical but the descriptive information and values were changed to avoid practice effects. A third pretest item was found to be problematic and was therefore replaced by an isomorphic item on the posttest. The order of presentation was modified in the posttest to further avoid practice effects (see Appendix E). Individual achievement was assessed by examining changes in performance from pretest to posttest based on scoring templates that were developed for evaluating the quality of students' responses (see Appendix F).

Homework Assignments

Students were given two homework assignments. The first assignment was given on the second day of the tutorial after students had been introduced to relevant statistical concepts and procedures. Students were required to answer a set of questions specific to a scenario presented in the assignment (see Appendix G). The questions provided students with an opportunity to explain the data collection process, to make comparisons, and to illustrate differences with a graph based on information presented in the problem.

The second assignment was administered after the knowledge acquisition phase (i.e., tutorial) and prior to the performance phase when the steps involved in conducting an experiment were explained. This assignment was more difficult than the first in that students were required to generate their own problem, explain their understanding of statistics in their own words, and make connections between concepts and procedures (see Appendix H). Scoring templates were developed for evaluating the quality of student responses and the extent to which students integrated their knowledge on each assignment (see Appendices I and J).

Assessments of Group Projects.

Presentations of group projects were scored on a 50-point scale based on criteria which reflect NCTM (1989) standards for "doing" mathematics and ASA (1991) guidelines for developing statistical content. These criteria include (a) the quality of the group's research question (5 points), (b) the procedures used for data collection (10 points), (c) the ways in which the data was presented to the class (10 points), (d) type of data analysis and quality of interpretation (10 points), (e) presentation style (10 points), and (f) creativity (5 points). These criteria were explained to students before they started their group projects and were used by groups to assess each presentation, including their own. Experimenters evaluating group projects were also trained on the six criteria. However, only four of the six experimenters could be present for group presentations. Thus, each group received four experimenter ratings, one self rating, and seven group ratings.

Presentations of Group Projects

Group presentations involved articulating research goals, results, interpretations, and conclusions to peers and mentors. Presentations were followed by question periods where the teacher, students, and experimenters asked groups for clarification or elaboration. Such question periods were included to (a) examine the understanding of students asking the questions as well as those providing the answers and (b) promote reflection.

Structured Group Journals

Structured journals were designed to be ongoing measures of group performance. Journals were structured in the sense that they required groups to answer specific questions (i.e., pr^mpts) about concepts, data, graphs, and project ideas. Such questions prompted students to formulate their knowledge, to reason about data and graphs, to plan their group projects, and to reflect on their learning (see Appendix K for prompts that were embedded within the group journals). The distribution of prompts within the journal was based on the nature of the activities groups were working on. Definition and explanation prompts (i.e., knowledge prompts) were located in the beginning to the middle of the journal since most of the instruction on concepts was given at this time. Reasoning prompts, on the other hand, were located in the middle to the end of the journal since groups worked mainly on databases and graphs. Planning prompts were found at the end of the journal since it was at this time that groups began to seriously think about their group projects. Reflection prompts, however, were evenly distributed throughout the journal.

Mini-Course Evaluation

A short questionnaire, consisting of eight questions, was included at the end of the posttest and administered on the last day of testing. Students were asked to rate how well they liked the course on a five-point scale. Open-ended questions allowed students to comment on the content of the course, to suggest improvements, and to reflect and comment on their group projects (see Appendix L).

CHAPTER 3: RESULTS

Two units of analysis were used in this study: individual subjects and groups of subjects. Quantitative and qualitative analyses were conducted to examine the amount and extent of individual and group learning. Quantitative analyses which examined individual learning included the following: (a) an Analysis of Variance (ANOVA) based on a split-plot design was conducted on pretest and posttest data to investigate whether the library of exemplars treatment was more effective in promoting statistical learning than the text treatment; (b) an ANOVA based on a between-factors design was performed on data from the first homework assignment to examine whether any differences existed prior to the assignment of treatments; and (c) descriptive statistics were used on data from the second homework assignment to investigate whether the library of exemplars treatment was more effective in fostering knowledge integration than the text treatment. Qualitative analysis of individual learning consisted of an examination of written responses to pretest and posttest as well as to both homework assignments.

Quantitative analyses which examined group learning included the following: (a) planned comparisons based on a split-plot design was conducted on project assessments to examine whether the library of exemplars treatment resulted in more similar and higher ratings of group projects than the text treatment and (b) a planned comparison based on a between-factors design was conducted on structured group journal data to examine whether the library of exemplars treatment was more effective in promoting planning than the text treatment. Qualitative analysis of group learning consisted of an examination of verbal data in presentations and written responses to structured journals as well as to course evaluations. Inter-rater reliabilities were conducted to examine the effectiveness of scoring templates on (a) pretest and posttest, (b) homework assignments, and (c) experimenter and group ratings of projects. This chapter reviews the results of quantitative and qualitative analyses for individual student learning first and then group learning.

Assessment of Individual Student Performance

This section reports results that were obtained using subject as the unit of analysis on two types of measures: (a) pretest and posttest and (b) homework assignments. The pretest and posttest measured students' statistical knowledge prior to and following instruction while the homework assignments measured students' knowledge during instruction. The results of quantitative and qualitative analyses are presented next.

Students' Statistical Knowledge Before and After Instruction

Total pretest and posttest scores served as indicators of students' statistical knowledge before and after instruction. A fifty point scale was used to score the tests. Since all test items consisted of open-ended questions, students were rewarded for correct explanations as well as for correct answers. A split-plot design using subjects as the unit of analysis was performed to examine whether or not type of treatment (text or library of exemplars) affected students' test scores (pre and post). Results from the Subject{Treatment (Library, Text)} x Test (Pre, Post) ANOVA demonstrated that there were no significant differences between the two treatments. However, there was a significant effect for test which indicated change in statistical knowledge for all students (see Table 2).

As Table 3 demonstrates, students in the text and library of exemplars treatments acquired a substantial amount of statistical knowledge as a result of the instruction. However, relative to the maximum test score students' overall performance was not exceptional. This finding suggested that although the instruction was effective in promoting students' statistical knowledge, it may not have been sufficient for acquiring depth of knowledge of all concepts. Perhaps the instruction was successful at fostering a deep understanding of only a few statistical concepts or a wide but superficial understanding of many concepts (i.e., breadth of knowledge).

Table 2

Subject(Treatment (Library, Text)) X Test (Pre, Post) Analysis of Variance

Source	Sum-of-Squares	*Df*	Mean-Square	*F*	*p*
Between					
Treatment	0.347	1	0.347	0.010	0.922
Error	566.792	16	35.425		
Within					
Test	590.130	1	590.130	50.311	**0.000***
Treatment X Test	10.797	1	10.797	0.920	0.352
Error	187.675	16	11.730		

*$p<0.05$.

Table 3

Mean Test Scores of Students in the Text and Library of Exemplars Treatment

	n	Pretest	Posttest
Text Treatment	11	6.636 (1.123)[a]	13.818 (1.739)
Library of Exemplars Treatment	7[b]	5.714 (1.408)	15.143 (2.180)

Note. Maximum score = 50.

[a]Mean (Standard Deviation)

[b]Three subjects were excluded from the analysis due to attrition.

Knowledge of specific concepts. In order to determine whether students' knowledge gains were limited to a few concepts or inclusive of all concepts, individual

ANOVAs based on a split-plot design were performed on test items measuring each of the following concepts: statistics, data, graph interpretation, outlier, hypothesis generation and identification, population, sample representativeness, sample size, randomization, sample, median, mean, and range. Since individual ANOVAs were conducted for each concept, the Bonferroni procedure (Kirk, 1982) was applied to adjust for Type I error in each analysis. The Subject{Treatment (2)} x Test (2) ANOVA demonstrated that there were significant test effects for the following statistical concepts: statistics, graph interpretation, hypothesis generation and identification, sample representativeness, sample size, sample, and range (see Table 4). No significant test effects were found for the concepts of data, outlier, population, randomization, mean, median, and mode. These results indicated that students' knowledge of statistical concepts, particularly those related to sampling, increased considerably as a result of instruction. This finding suggested that students acquired knowledge of _many_ concepts rather than a few.

In addition to testing the main effect of test on each item, the Subject{Treatment (Library, Text)} x Test (Pre, Post) ANOVAs explored the interaction of treatment by test. Significance was determined by the Bonferroni procedure for adjusting type I error. The results indicated that there was a significant interaction between test and treatment for three concepts: sample representativeness ($F(1,16)= 8.581, p<0.01$), sample size ($F(1,16)=6.862, p<0.01$), and statistics ($F(1,16)= 4.899, p<0.01$). Figure 8 illustrates the cross-over interaction between test and treatment for sample representativeness. The graph suggested that students in the library of exemplars treatment outperformed students in the text treatment, particularly at posttest. Figure 9 also suggested better improvement on sample size items for students in the library of exemplars treatment. However, Figure 10 demonstrated that differences between treatments on an item measuring students' understanding of the purpose of statistics seemed greatest for students in the text treatment at pretest.

Table 4

Test Performance on Items Measuring Statistical Concepts and Procedures

Concept or Procedure (pre item #; post item#)	Maximum Score	Pretest Mean (*SD*)	Posttest Mean (*SD*)	*F*	*p*
Statistics (2; 4)	2	0.500 (0.514)	0.944 (0.236)	12.214	0.003*
Data (9b; 2b)	2	0.889 (1.023)	1.389 (0.850)	6.701	0.020
Graph Interpretation (4a; 9a)	2	0.278 (0.575)	0.944 (0.539)	14.473	0.002*
Outlier (8; 6)	2	0.000 (0.000)	0.500 (0.786)	6.862	0.019
Hypothesis Generation and Identification (1a, 6a; 5a, 7a)	4	1.556 (1.247)	2.889 (0.963)	14.141	0.002*
Population (6b; 5b)	3	0.278 (0.575)	0.833 (0.514)	7.566	0.014
Sampling					
Sample Representativeness (1d, 4b, 6d, 10b; 7d, 9b, 5d, 8b)	4	0.556 (0.922)	1.778 (1.263)	22.712	0.000*
Sample Size (1c, 4b, 6d, 10b; 7c, 9b, 5d, 8b)	5	0.389 (0.608)	1.333 (0.767)	46.622	0.000*
Randomization (9a; 2a)	2	0.167 (0.383)	0.444 (0.616)	2.372	0.143
Sample (1b, 6c; 7b, 5c)	4	0.833 (0.786)	1.444 (0.511)	12.287	0.003*

Table 4 cont'd....

Concept or Procedure (pre item #; post item#)	Maximum Score	Pretest Mean (*SD*)	Posttest Mean (*SD*)	*F*	*p*
Measures of Central Tendency					
Mode (5a, 5d; 1a, 1d)	3	0.000 (0.000)	0.000 (0.000)	-	-
Median (5b, 5d, 7; 1b, 1d, 10)	5; 4[a]	0.000 (0.000)	0.278 (0.752)	1.827	0.195
Mean (5c, 5d, 7; 1c, 1d, 10)	4; 5	0.389 (0.502)	1.000 (1.188)	3.616	0.075
Measures of Variation					
Range (3, 10a; 3, 8a)	4	0.056 (0.236)	0.833 (0.985)	10.670	0.005*
Total Score:	46[b]				

Note. The dash (-) indicates that no analysis was conducted for mode due to the lack of variation in the dependent variable.

[a] The difference in maximum pretest and posttest scores (;) for the concepts of median and mean were due to the different solutions for each (see Appendix F).

[b] The maximum score on the pretest and posttest is 50, however, the total score relating specifically to explanations of the concepts listed in this table is 46.

**p*<0.01 based on the Bonferronni adustment for Type I error.

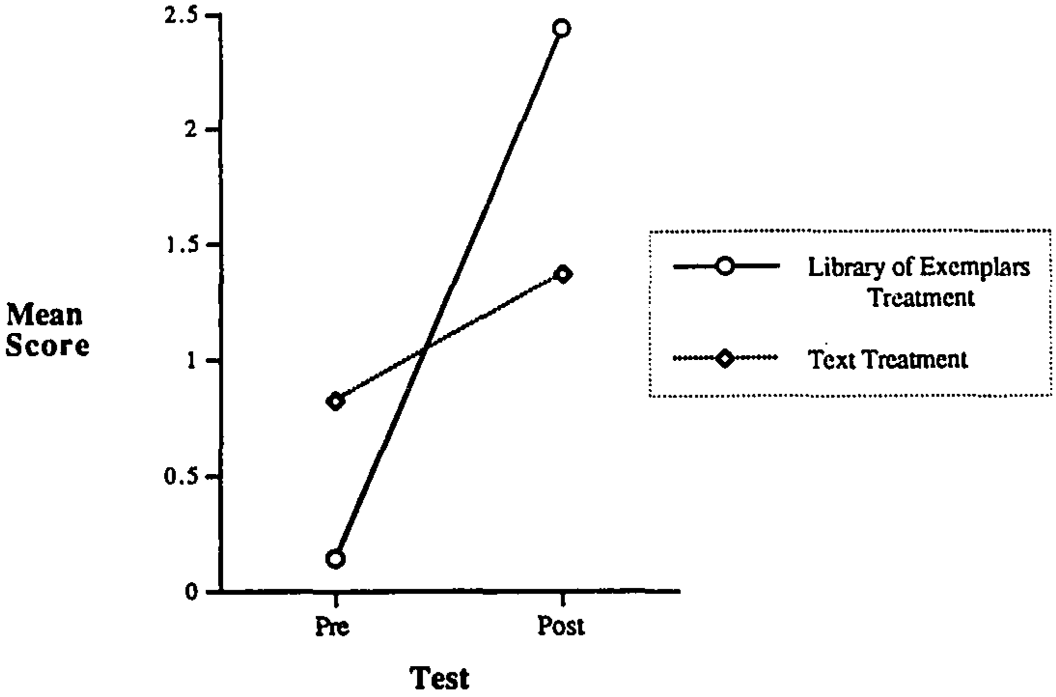


Figure 8. Interaction of condition and test on the concept of sample representativeness.

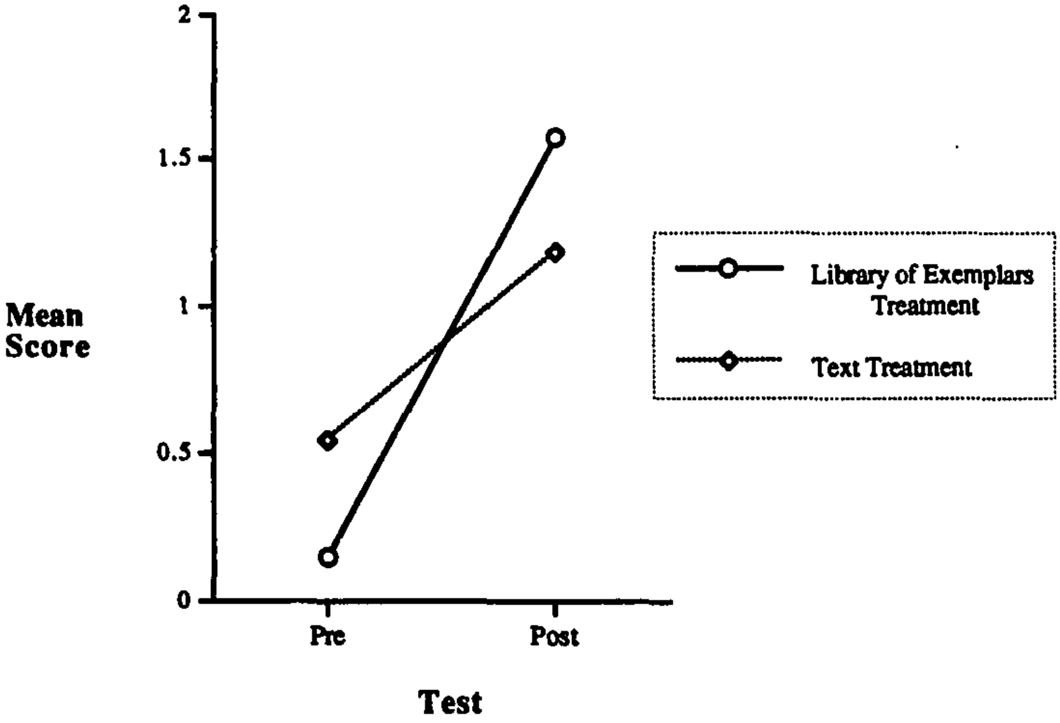


Figure 9. Interaction of condition and test on the concept of sample size.

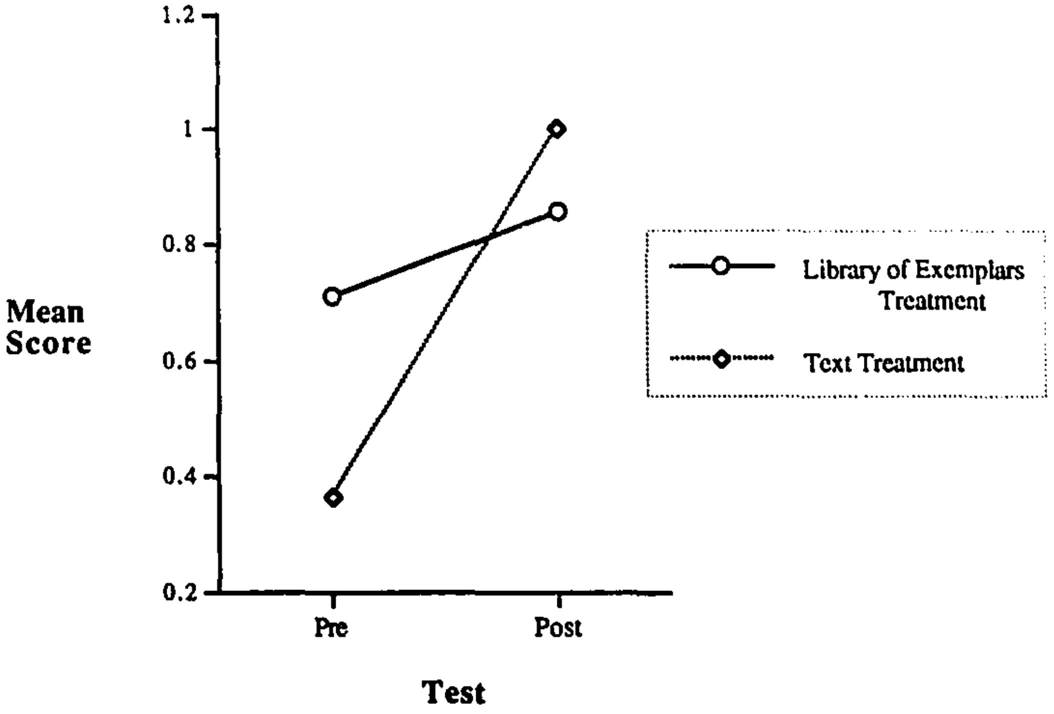


Figure 10. Interaction of treatment and test for the concept of statistics (purpose of).

Post-hoc analyses were performed to examine the nature of the interactions for the concepts of sample representativeness, sample size, and statistics. The Scheffé S procedure (Kirk, 1982, pp. 121-122) for making comparisons was used given the unequal n's in each cell (i.e., *n*=11 for text and *n*=7 for library of exemplars). Results indicated that the difference in performance from pretest to posttest on items of sample representativeness was significant for the library of exemplars treatment (*F*(1, 16)= 24.225, *p*<0.05) but not for the text treatment (*F*(1, 16)= 2.172, *p*>0.05). Moreover, this difference was significant at posttest (*F*(1, 16)=6.425, *p*<0.05) but not at pretest (*F*(1, 16)=2.581, *p*>0.05). Results for sample size indicated that there was a significant pretest and posttest difference for both the library of exemplars (*F*(1, 16)=36.413, *p*<0.05) and text (*F*(1, 16)= 11.388, *p*<0.05) treatments. Finally, results for the concept of statistics indicated that there was a significant difference between the library of exemplars and text treatments at pretest (*F*(1, 16)= 4.944, $p<0.05$) but not at posttest ($F(1, 16)=0.825, p>0.05$) and that the difference from pretest to posttest was significant for the text treatment ($F(1, 16)=20.988, p<0.05$) but not for the library of exemplars treatment ($F(1, 16)=0.675, p>0.05$).

Although the analyses of test performance on each concept indicated that students understood most of the concepts, the nature of this knowledge was not revealed by the analysis. In order to determine the extent of students' prior knowledge and whether, as a result of instruction, students acquired depth or breadth of knowledge, responses to test items (pre and post) were analyzed qualitatively.

Qualitative analysis of responses to pretest items indicated that prior to instruction students had (a) knowledge of statistical concepts such as data and statistics, (b) experience with column graphs, and (c) experience with problem scenarios or case-based problems that provided relevant information for answering questions (e.g., Ben Johnson incident used as a way of introducing the problem of steroid use and type of athletics). Students' everyday understanding of "data" was reflected in their definition: Data is "information." Students also seemed to have an idea of the purpose of "statistics" as demonstrated in the following statements: "If you wanted to inform the public of a problem... if you wanted to find out how the population regarded your business or anytime you want to find out what the population's opinion or status is," "Once I got all the information I would use the statistics," "I would use statistics when I was looking for a certain amount of something," and " When you would want to see if a player in a sport like hockey, is good enough to play on a team." These definitions reflect an everyday understanding of statistics which may have been acquired through exposure from the media such as newspapers.

Students' previous experience with graphs and problem scenarios was similarly demonstrated in their responses to pretest items. Three of the five pretest items attempted by students were those that presented a graph (i.e., popularity of two radio morning shows) or a problem scenario (i.e., favorite fastfood restaurant and steroid use in athletics--the Ben Johnson incident). Students' familiarity with graphs was also demonstrated in a class discussion initiated by the teacher in which newspaper clippings were used to prompt students to communicate their knowledge of graphs. The teacher subsequently used this information to introduce statistics. The following dialogue is an example of how the concept of data was introduced based on a discussion of graphs.

Teacher: What's another kind of graph or chart?
Student1: Line graph
Teacher: Line! Just using lines.
(to the class). Have you ever used one like that? Have you ever done any lines?
Student2: Yeah.
Teacher: Where? And where have you used a line?
Student3: Science.
Teacher: In science. To show what?
Student4: To show the object of the graph, to show the results.
Student5: To show um...(draws an x and y axis with his hand in the air so everyone can see).

Teacher: All these charts or graphs are based on what though, first of all? What do you have, what do you collect?
Student6: Answers a certain question.
Teacher: Yes, answers certain questions. And what is it that you collect, what do you call that?
Student7: Data
(Other students agree with Student7's response)
Teacher Data? Another word for data is what? What is data, that's a very simple word for ah...
Student3: Information
Teacher: Information, yeah. Anything else? Data? What are you ah, tallying up? What do you call that?
Student3: The results
Teacher: In the form of... words... or... letters... or...?
Student8: Numbers.

The nature of students' responses to posttest items referring to the concept of data indicated that students' understanding was limited to their everyday definition. Students' level of understanding of the concept statistics and ability to interpret graphs was basic despite an increase in knowledge. For instance, students tended to limit their descriptions of statistics to activities they did in the course--conducting surveys. However, the specific purpose of surveys (i.e., to describe or make predictions) was never mentioned. The tendency to describe or define concepts based on individual experience was also evident in students' interpretation of graphs. Conclusions about information presented in the graph were usually based on students' own assumptions and high school experience. For example, one of the postiest items (see item #9 in Appendix D) used a column graph to illustrate the number of hours grade 8 students studied compared to the number of hours grade 11 students studied. Rather than explaining how they arrived at their conclusion, learners explained why they thought the result was valid: Grade 11 students studied more than grade 8 students because they had more (a) work to do, (b) difficult tests and homework, and (c) pressure because they wanted to graduate. These comments suggested that learners tried to make sense of the graph through what they already knew of the real world; i.e., the higher the level of education the more hours of study required. However, students tended to limit their answers to the general case while neglecting the specifics of the problem such as the increased number of hours studied each semester. Nonetheless, the congruence of the information with students' assumptions seemed to have helped learners' interpret the graph.

Qualitative examination of responses to the remaining pretest items suggested that students had no prior knowledge of concepts and procedures such as mean, median, mode, outlier, population, randomization, range, and sampling. Students' lack of prior knowledge of concepts such as mean, median, mode, outlier, and randomization was demonstrated in their inability to answer pretest questions that probed this knowledge. Posttest responses indicated that despite instruction, students still had difficulty with items pertaining to measures of central tendency, oulier, and randomization. For instance, on both the pretest and posttest, students either omitted questions that required calculating the mode and median or gave inaccurate answers. Students also had difficulty distinguishing the mean from the median, often defining the median as the "average" rather than the "middle number," terms which in the instruction, were used exclusively to define the mean and median respectively. In addition, students were unable to construct a graph or chart that demonstrated an outlier despite instruction. Student difficulties with the concept of randomization were not as obvious. Despite attempts to define randomization on the posttest, students were not always explicit in their description of the randomization process. For example, most students described randomization as "asking different people" or "randomly picking a group out of a group." Such responses suggested that students knew that randomization was related to sampling, however, they had difficulty expressing exactly what the term meant. Only a few students were more explicit and explained for instance, that randomization was "when something is picked in no particular order. For example, picking names out of a hat to develop teams... everybody has an equal chance so its fair." Finally, students' understanding of population was restricted to an everyday definition such as "a large group."

Although students lacked prior knowledge of the concept range, responses on posttest items indicated that they had acquired sufficient understanding of how to calculate the range. Students' knowledge of sampling prior to instruction was restricted to an understanding that a sample was a subset or example of something. However, by the end of the instruction, students seemed to have a good grasp of the sampling process although the generality of their answers suggested that depth of knowledge had not been acquired. In general, it seemed that students' knowledge remained at a superficial level and that breadth rather than depth of knowledge had been acquired.

Influence of group collaboration on individual performance. Although the role of group collaboration on individual performance was not systematically examined in this study, an informal analysis was conducted. The question of interest was whether students' scores were more similar to those of their group members on the posttest than they were on the pretest. Total pretest and posttest scores of students within each group were compared by eyeballing the data to answer this question. Results indicated that students' scores within four of the six groups did not differ substantially from each other which suggested that their statistical knowledge prior to instruction was about the same (see Table 5)[3]. This consistency was maintained after instruction in most of the groups which suggested that students within each group learned at approximately the same rate. Students of all ability levels demonstrated substantial knowledge gains from pretest to posttest, however, some of the greatest gains were made by high ability students (see Table 5). These findings suggested that group collaboration may be beneficial for learning.

Inter-rater reliability. Inter-rater reliabilities were conducted on the pretest and posttest. Two graduate students scored both tests using templates that were developed for each (see Appendix F). The Pearson Correlation Coefficient indicated a high overall correlation (r=0.991) between raters. Inter-rater reliabilities were also strong for the pretest (r=0.982) and posttest (r=0.987). The strong correlation between raters suggested that the scoring criteria were clear.

Students' Statistical Knowledge During Instruction

Total homework assignment scores served as indicators of students' knowledge during instruction. Although several other ongoing measures such as computer screen recordings (i.e., visual recordings of students' work on the computer using Mediatracks software), audiotapes, and videotapes were used as part of a larger study, the analysis in this manuscript is limited to data from the homework assignments. A ten-point scale which rewarded students for correct answers and correct explanations was used to score the two assignments. Level of explanation on a well-defined problem and knowledge integration on an ill-defined problem determined the level of performance on the first and second assignments respectively (see Appendices I and J). The data for both assignments was analyzed quantitatively and qualitatively.

[3] Six groups were compared rather than eight because of attrition during the posttest. The absence of two students in a group of three and of one student in a group of two eliminated the possibility of examining within group differences.

Table 5

Within Group Test Performance

Treatment	Group	Student's Ability Level	Pretest Score	Posttest Score	Improvement (Post - Pre)
Text					
	Group A				
		Medium	5	12	7
		High	13	16	3
		Medium	9	16	7
	Group B				
		Medium	7	14	7
		High	10	28	8
	Group C				
		Medium	5	10	5
		Low	3	11	8
		High	14	13	-1
	Group D				
		High	0	12	12
		Medium	4	13	9
		Low	3	7	4
Library of Exemplars					
	Group E				
		High	9	21	12
		Low	5	13	8
	Group F				
		Medium	6	13	7
		High	8	16	8
		Low	4	15	9

Assignment 1: Statistical knowledge in a well-defined problem. Students' statistical knowledge at the time the first homework assignment was administered was expected to be somewhat limited. Descriptive statistics which were used to examine overall results indicated that students' understanding of statistics and of the data collection process was about average (*M*=5.063, *SD*=1.652, *n*=15). Qualitative analysis of written responses was conducted to examine the nature of students' understanding. Results indicated that pupils were able to identify hypotheses and draw graphs to illustrate results, however, they were unable to go beyond the problem statement to (a) generate their own explanation of the data collection process, (b) compute the mean and range as a basis for making comparisons, and (c) draw clear conclusions. For example, when asked to explain how a researcher, Dr. Aloe, gathered his data given that he randomly assigned twenty grade 6 mathematics students to a computer and non computer group, students tended to recite information directly from the problem statement rather than explain the data collection process in their own words. Students used the term "random" without elaborating on its purpose in Dr. Aloe's study.

Difficulties going beyond the problem statement were also illustrated when students were required to demonstrate how they would compute differences based on data provided in the problem (i.e., mathematics scores for the computer and non computer groups). Most students did not seem to know what to do with these scores, either omitting the item or taking the question literally and subtracting rows of numbers across variables. Conclusions were consequently vague and ambiguous. Group differences were attributed to the assignment of groups to two treatments rather than to differences in performance scores which would have been highlighted by the mean and range. The finding that students were unable to extend beyond the problem statement was not surprising given that assignment one was administered immediately after the introductory lecture when knowledge had just been acquired.

An ANOVA with a between-groups design was conducted to determine whether differences between students existed prior to receiving the text and library of exemplars treatments. Results indicated that differences between treatments was not significant ($F(1, 14)=0.195, p>0.05$).

Assignment 2: Integration of statistical knowledge in an ill-defined problem. Analysis of data for the second homework assignment was restricted to the use of descriptive statistics due to a poor response rate (i.e., 38.1%). Overall performance and treatment differences were examined. Analysis of overall performance suggested that students had difficulty integrating their knowledge of statistical concepts and experimental procedures ($M= 4.563$, $SD=1.700$, $n=8$). Most students listed or defined concepts without explaining how these related to the experimentation process. Some students, however, were able to make this connection, as illustrated in the following example:

> When doing an experiment, you must first have a hypothesis and then find the population you want to ask your question to. After doing this, select your sample group and then develop a method of collecting data as in a type of graph. After doing this, you may analyze it by finding the mean (average), the median (the middle number), the mode (the most popular number), and the range (difference between #s).

Although this excerpt is an example of how one student was able to make some connections between concepts and procedures, most students failed to do so. Difficulties with assignment two, however, may not be solely attributed to students' inability to integrate their knowledge. The increased level of difficulty from the first to the second assignment resulted in a dramatic drop in response rate, from 76.2% in assignment one to 38.1% in assignment two, and in a slight decrease in performance, from $M=5.063$ to $M=4.563$ respectively. Since both assignments examined the same concepts and procedures and the second assignment was more ambiguous and ill-defined than the first, the poor response rate and average performance of students on assignment two may also be due to the nature of the task itself and not strictly to an inability to make connections. The second homework assignment required that students generate a problem and explain statistics in their own words rather than finding part of the answer in the problem statement. This requirement made students anxious. As a result, assignment two had to be explained several times. Yet despite these explanations, the confusion persisted. For instance, rather than explaining the four steps involved in conducting a study, some students listed tasks they did as participants in this study (i.e., reading and working with the tutorial, writing in the structured group journals, etc.). Whether students misunderstood the task or did not know the answer was difficult to determine given the low response rate.

Analysis of treatment differences examined the influence of treatment (text or library of exemplars) on performance. Since the assignment was given after groups had been assigned to the two treatments, students receiving the library of exemplars were expected to outperform those receiving the text. However, given that only eight students returned the assignment, the significance of this hypothesis could not be tested. The analysis was therefore restricted to the use of descriptive statistics. Results suggested that the library of exemplars approach (*M*=5.500, *SD*=1.323, *n*=3) was more effective at helping students make connections between statistical concepts and procedures than the text approach (*M*=4.000, *SD*=1.768, *n*=5). The significance of these results, however, is undetermined.

Influence of group collaboration on individual performance. Of interest was whether group collaboration in solving problems affected the performance of individual students within each group and whether this influence differed according to ability levels. Scores on the first assignment demonstrated various levels of performance within each group, each in accordance with ability levels: High ability students scored the highest while low ability students scored the lowest. These observations suggested that the influence of group problem solving on the performance of individual students was minimal on assignment one. Given that students had not previously worked together in groups and that this assignment was administered early in the instruction, this result was not surprising. Students had to adjust to working with peers who were not necessarily friends and learn to solve problems collaboratively. At this point in the instruction, students spent most of their time adjusting to rather than learning from each other. However, it was expected that once adjustments were made, all students, particularly low ability students would benefit from group problem solving. Unfortunately, the low response rate did not allow for the examination of individual differences within any of the groups. As a result, comparisons could not be made to determine whether group problem solving influenced the performance of individual students on the homework assignments.

Inter-rater reliability. Inter-rater reliabilities were conducted on both homework assignments. Two graduate students used templates developed for scoring assignment one (see Appendix I) and assignment two (see Appendix J). The Pearson Correlation Coefficient indicated high correlation between raters on assignment one (*r*=0.874) and assignment two (*r*=0.946). Closer examination of ratings for assignment one suggested that the lower correlation was due to the ambiguity of one of the items.

Assessment of Group Performance

Previous analyses of student performance on the pretest and posttest as well as on the homework assignments were based on subjects as the unit of analysis. Each student received his or her own score for each of these measures. However, other measures, such as group projects and journal entries, assessed group rather than individual performance. The next section reports quantitative and qualitative results for group projects and journals based on group as the unit of analysis.

Assessments of Group Projects

The effectiveness of the library of exemplars approach to assessment was compared with the text approach based on assessments of group projects. Such assessments provided quantitative data for measuring group performance. The overall score on the six project assessment criteria (i.e., maximum of 50) was the dependent measure. Three types of raters assessed group projects: group self-assessments, group assessments of other groups, and experimenter assessments of all groups. These raters represented three levels of the independent variable assessment type. Since each presenting group received one self-assessment score, 7 group scores, and 4 experimenter scores, ratings for each assessment type were averaged for the purpose of analysis. Planned comparisons were conducted to test the following hypotheses: (a) the library of exemplars approach would provide more consistency between student and experimenter assessments than the text approach given the explicitness of the assessment criteria in the former treatment; and (b) ratings in the library of exemplars treatment would be higher than ratings in the text treatment.

The hypothesis that the library of exemplars approach would provide more consistency between student and experimenter assessments than the text approach was tested using two one-tailed planned comparisons. The contrast compared experimenter and student assessments (i.e., self and group assessments) in the text treatment with experimenter and student ratings in the library of exemplars treatment. Note that group scores in this analysis were based on ratings that groups in each condition *gave* other groups rather than the scores each received since the hypothesis dealt with differences in assessments given by each type of rater. Results indicated that there were no differences between treatments on ratings given by experimenters and students (F(1, 6)=5.405, p>0.10). These findings did not confirm the hypothesis that the library of exemplars approach would provide more consistent ratings of projects than the text approach. However, overall mean scores of groups assessing themselves (M=45.25) and other groups (M=39.55) indicated that groups were more stringent when assessing other groups' performance than they were assessing their own projects. These values suggested that groups were motivated by competition.

The hypothesis that ratings in the library of exemplars treatment would be higher than ratings in the text treatment was tested using a one-tailed planned comparison. Contrary to the previous analysis, group ratings in this analysis were based on scores received by each group. Results indicated that treatment differences were not significant ($F(1, 6)= 0.038$, $p>0.10$). Mean ratings in the text (M=40.03) treatment were similar to those in the library of exemplars (M=42.04) treatment. However, the mean overall score (M=41.07) indicated that on the whole, groups performed well. These results suggested that students, irrespective of treatment, were able to adapt their performance accordingly as well as to assess their own performance and those of others.

Inter-rater reliability. Pearson product-moment correlations were conducted to examine inter-rater reliabilities between (a) experimenters' overall ratings, (b) experimenters' ratings on each criterion, and (c) groups' overall ratings. Missing data in experimenter ratings of groups in the text treatment (i.e., overall and on each criterion) were substituted by the mean of ratings in that treatment to perform the inter-rater reliabilities. Results of the overall inter-rater reliabilities between experimenters indicated that experimenters 1 and 2 (r=0.626), 1 and 4, (r=0.657), and 2 and 3 (r=0.586) were moderately correlated while all other combinations were low (see Table 6).

Table 6

Overall Inter-Rater Reliabilities Between Experimenters

	Expter 1	Expter 2	Expter 3	Expter 4
Experimenter 1	1.000			
Experimenter 2	0.626	1.000		
Experimenter 3	0.174	0.657	1.000	
Experimenter 4	0.586	0.136	-0.364	1.000

The correlations presented in Table 6 indicate the consistency of experimenters' ratings of projects across the six criteria, however, they do not specify whether inconsistencies were due to difficulties in rating every criteria or one criterion in particular. In order to determine whether some criteria were more problematic for experimenters than others, inter-rater reliabilities between experimenter ratings on each criterion were conducted. However, given the allocation of points on each criterion (5 or 10) and the small sample size, an analysis based on *agreement* was deemed more meaningful than one based on correlations. Agreement was determined based on the assumption that a difference of 1 point between experimenter ratings on any criterion is not significant. The following rule for calculating agreement was therefore devised: If the absolute difference between two experimenters' ratings on any criterion is 1, then the two experimenters agree but if this absolute difference exceeds one, then the experimenters disagree. The frequency of agreement was then used to determine whether experimenters' ratings on each criterion were relatively consistent. Table 7 demonstrates the frequency of agreement between experimenters for each treatment (i.e., library of exemplars and text) and across treatments (i.e., total). For example, a frequency value of 3 between experimenters 1 and 2 rating groups in the text treatment on the quality of question criterion means that the ratings given by these experimenters were the same for 3 of the 4 groups in this treatment (i.e., difference between scores did not exceed an absolute value of 1).

The data presented in Table 7 indicated that overall, the agreement between most experimenters was low or moderate (i.e., at least half of the frequency values were 6 or less) or inconsistent (i.e, large range in values) on the data collection, data analysis, and presentation style criteria. In contrast, the criteria that seemed the least problematic for experimenters to score were quality of question and creativity. Data presentation on the other hand, seemed somewhat problematic. The data also suggested that there was high

Table 7

Agreement Between Experimenters on Each Project Criterion

Criterion	Expters (*n*=4)	Library of Exemplars (*n*=4)	Text (*n*=4)	Total (Overall Agreement[a])
Quality of Question	1 & 2	4	3	7
	1 & 3	4	4	8
	1 & 4	4	3	7
	2 & 3	4	4	8
	2 & 4	4	2	6
	3 & 4	4	3	7
Data Collection	1 & 2	4	2	6
	1 & 3	3	3	6
	1 & 4	2	2	4
	2 & 3	3	3	6
	2 & 4	1	0	1
	3 & 4	2	0	2
Data Presentation	1 & 2	3	3	6
	1 & 3	4	3	7
	1 & 4	3	2	5
	2 & 3	3	4	7
	2 & 4	4	2	6
	3 & 4	4	2	6

Table 7 cont'd...

Criterion	Expters (*n*=4)	Library of Exemplars (*n*=4)	Text (*n*=4)	Total (Overall Agreement[a])
Data Analysis	1 & 2	3	2	5
	1 & 3	3	3	6
	1 & 4	4	3	7
	2 & 3	1	3	4
	2 & 4	4	3	7
	3 & 4	3	2	5
Presentation Style	1 & 2	3	1	4
	1 & 3	2	3	5
	1 & 4	2	2	4
	2 & 3	4	4	8
	2 & 4	3	2	5
	3 & 4	3	2	5
Creativity	1 & 2	4	3	7
	1 & 3	4	4	8
	1 & 4	4	3	7
	2 & 3	4	4	8
	2 & 4	4	2	6
	3 & 4	4	3	7

[a] Overall Agreement = Experimenters gave the same score to *n* of the 8 groups on the criterion in question (i.e., Library of Exemplars + Text).

agreement between experimenters when rating groups receiving the library of exemplars treatment on the following criteria: quality of question, data presentation, data analysis, and creativity. Consistency in agreement was also strong for experimenters scoring groups in the library exemplars condition. Conversely, the ratings of the same experimenters were in high agreement for the quality of question and creativity criterion. Consistency in agreement was not as high as for the library of exemplars treatment. These results suggested that the low overall inter-rater reliabilities between experimenters was due to problems in scoring performance on the data collection, data analysis, and presentation style criteria. In addition, the data suggested that there was less agreement and consistency in agreement between experimenters scoring groups in the text treatment (i.e., data collection, data presentation, data analysis, and presentation style) than in the library of exemplars (i.e., data collection and presentation style) on most of the criteria.

Given that groups of students as well as experimenters were given instruction on how to score group projects, inter-rater reliabilities of the groups' ratings were performed in addition to experimenter assessments. The results of this analysis indicated that correlations between groups in their rating of projects were moderately low, ranging from *r*=0.011 to *r*=0.627 (see Table 8). In addition, the correlations between groups in the library of exemplars treatment were generally higher than between groups in the text treatment, although the correlations in both were rather low. An interesting observation was that most of the strongest correlations were between the first group in the text treatment with each of the first three groups in the library of exemplars condition (*r*=0.551, 0.568, 0.542 respectively). These findings suggested that consistency of ratings given by groups were moderately low for both treatments.

Table 8

Inter-Rater Relibilities Between Groups

	Textl	Text2	Text3	Text4	Libraryl	Library2	Library3	Library4
Textl	1.000							
Text2	-0.519	1.000						
Text3	0.161	0.011	1.000					
Text4	0.081	0.129	0.222	1.000				
Libraryl	0.551	0.242	-0.157	-0.248	1.000			
Library2	0.568	-0.399	0.144	-0.275	0.254	1.000		
Library3	0.542	-0.321	0.129	0.454	0.159	-0.355	1.000	
Library4	-0.268	0.027	0.627	0.196	-0.587	0.346	-0.506	1.000

Presentations of Group Projects

Videotape recordings of group presentations provided the data for examining group performance qualitatively. Of interest were the types of projects groups developed and students' ability to communicate their knowledge during the presentation and subsequently during the question period. Table 9 summarizes the information conveyed in group presentations by illustrating the types of research questions that were generated by each group, where the data was collected, the types of data collected, and the types of analyses and graphs that were conducted. Groups' preference for conducting surveys was evident from their research questions. Most groups chose topics which reflected their interests, such as sports and music, while others were more interested in issues which affected them directly, such as the installment of condom machines in high schools. In trying to answer their question, groups generally surveyed classmates who were also participating in this study. The mathematics class provided a pool of subjects that was immediately accessible. Other groups decided to survey additional people and thus included students who were not in their mathematics class in their sample. Only one group did not conduct a survey for their mini-experiment. Rather, Group D (see Table 9) used hockey data from the 1989 season already at their disposal and supplemented this with data for subsequent seasons which they collected from city newspapers.

Given students' group projects, it was not surprising that the type of data collected was frequency data . However, the nature of such data posed some problems for Group B (see Table 9) when doing their analyses. Although students in the group were eager to apply and demonstrate their new knowledge and did so by identifying outliers and calculating the mean, mode, median, and range; they neglected to discriminate among these measures in terms of their appropriateness for the problem. In other words, the group blindly used all possible measures rather than selecting the most appropriate one for their data. Other groups calculated percentages rather than means for analyzing their data. Perhaps this reflected an understanding that the latter measure was not ideal for analyzing frequency

Table 9

Summary of Group Presentations

Group[a]	Research Question	Data Source	Data Type	Analyses	Graphs
A	Favorite school subject between English, French, math, history, M.R.E, &gym	Mathematics class	Frequencies	Percentages	Pie chart &Bar graph
B	Preferred airline between Air Canada, Delta, Canadian Airlines, and American Airlines	Mathematics class	Frequencies	Mean, median, mode, & range.	Bar graph
C	Favorite fastfood restaurant between Harveys, Pizza Hutt, McDonald's, Wendies, & Lafleur's	Mathematics class	Frequencies	Percentages	Column graph
D	The Montreal Canadien's record for the last 3 years	Newspaper & datafile	Hockey Scores	-	Column graph
E	Favorite basketball team between the Chicago Bulls, New York Nicks, Portland Trailblazers, Orlando Magics, Boston Celtics, Pheonix Suns, & San Antonio Spurs	Mathematics class & students from 2 other classes	Frequencies	Percentages	Column graph & Pie chart
F	Stores people go into the most and which stores have the best prices	Mathematics class & other grade 8 students	Frequencies Frequencies	- Percentages	Bar graph & Pie chart
G	Whether grade 7 and 8 students think LPHS should provide condom machines in the washrooms	Grade 7 and grade 8 students at high school	Frequencies	Percentages	Column graph & Pie chart
H	Favorite music groups between TLC, Kiss, Bon Jovi, Guns 'n Roses, Eric Clapton, Snow, & Metallica	Mathematics class & other grade 8 students	Frequencies	Percentages	Pie chart

[a]All groups listed in this table refer to the groups listed in Table 5.

data, however, responses during the question periods seemed to suggest that graphs were deemed sufficient for conveying the results.

Two types of graphs were used by groups to demonstrate their findings: a column or bar graph for illustrating frequency data and a pie chart for depicting percentages. Students in Group C (see Table 9), however, seemed confused about the relationship between data and graphs. A column graph with frequency data was used to discuss results of the group's survey on favorite fastfood restaurants (see Figure 11), however, the results were discussed in terms of percentages. Classmates watching the presentation became confused at the incongruency of the information depicted in the graph and what was being said by the group. The following dialogue between presenters (C1, C2, and C3), classmates (Student 1 and Student 2), and teacher illustrated this confusion.

Teacher: How did you know there was a certain percent? Is there a label on the chart? Um... that (graph) says the numbers along the bottom are percentages... Is that what it means?
You were saying... I was trying to follow you... you were saying percent but I don't see percentage signs or labels of percent. I was wondering uh... were you giving that from memory or reading off the chart?

C1: From memory

Student1: So, I think that what they're trying to say is that the number that they took... I think that the number of students they asked... not a percent. They divided the whole number of students they asked and said it was a percent. I guess they should have used a pie graph.

Teacher: Okay. Well, you said that McDonald's was 40% and Harveys was how many?

C1: No 50%

Teacher: Yes, well um... Wendy's

C1: 20

Teacher: Well yeah, you're just reading numbers off there... So you have a label along the bottom of the graph, along the X-axis... uh... that says "number" ... does it really represent percent?

C1: Uh...

Teacher: Because there was a relationship between what you are saying and

	the numbers that are at the bottom of each of those bars there.
C2:	Um, yeah... Um... Yeah, I guess so.

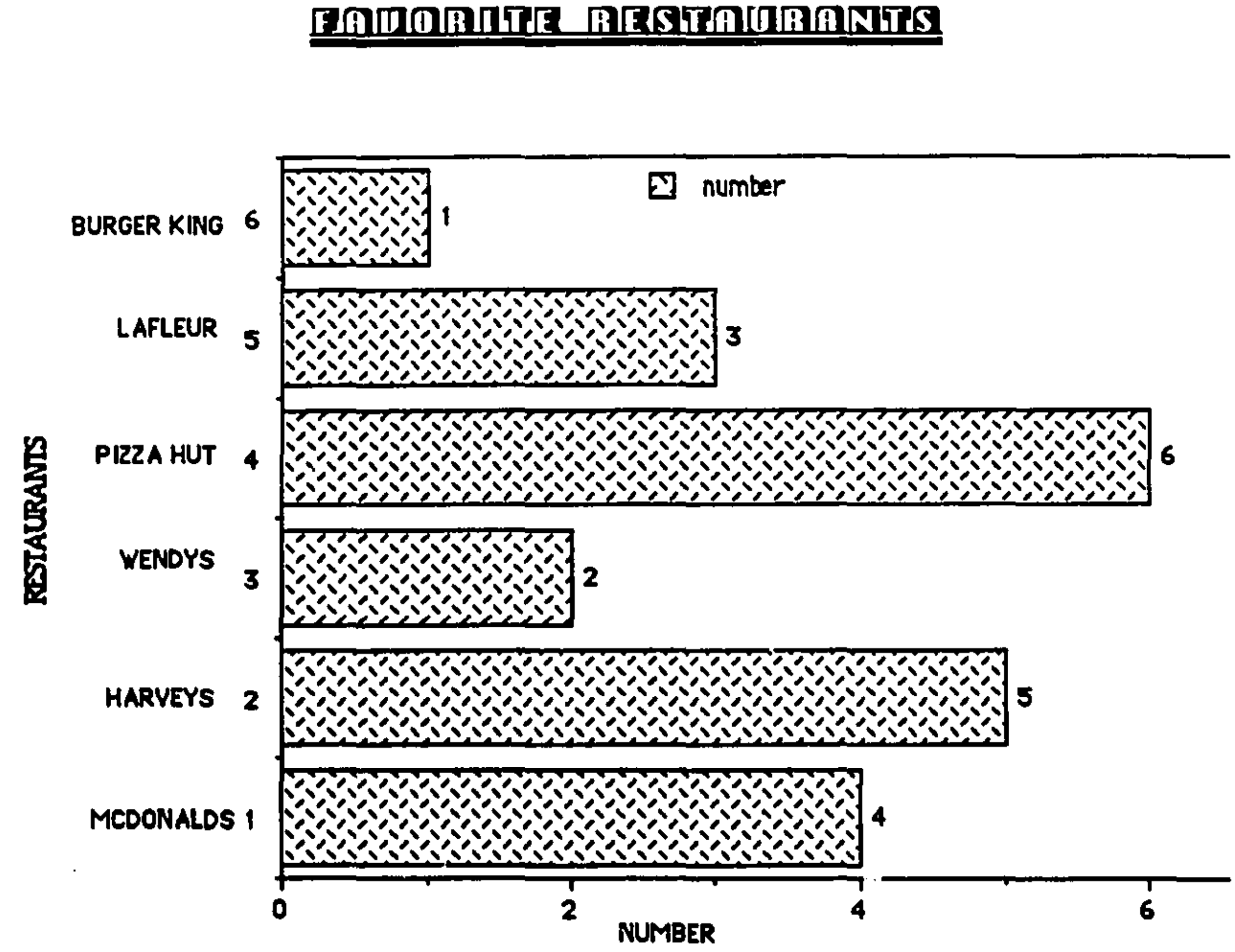


Figure 11. Group C's graph.

As the discussion continued, it became clear that Group C had calculated the percentages by hand and that errors had been made in the process.

Teacher:	Well then, 4 out of 21 is not 40%
C1:	Show him what you got (to C2)
Teacher:	If 4 people chose McDonald's ... is that what you're showing here? Then 4 out of 21 is not 40%
C1:	Eh, C2...
Teacher:	The numbers here can't be percent because they don't make 21, see what I mean?

C1: C2...
C3: Well, what I have here are the original positions that we have... uh on our graph there. Ah... for Burger King there is 19%.
Teacher: Yeah...
C3: And for Lafleur's 24%
Teacher: That's not what C1 was saying though.
C3: Pizza Hut is 10%... ah... Wendy's is 29%, Harveys is 14 and McDonald's is 4.
Student2: Excuse me C3, Pizza Hut can't be 10, especially when Wendy's is 50%... or whatever you said... 25.

None of the students in the group had verified the accuracy of these calculations nor were they aware of how these had been made as demonstrated in this quote: "C2 is the one who was doing calculations for percent...not me and C1 didn't do anything!" Group cohesiveness broke down under pressure and students failed to take responsibility for the work they had done as a group. In the end, Group C was unable to fully answer the teacher's questions or those of their classmates.

Rather than generating one research question and using one or two graphs to depict the results, some groups generated two questions and created a different graph for answering each question. Group F, for example, used a bar graph for illustrating their results on the types of stores people went to the most while a pie chart was used to demonstrate which stores were thought to have the best prices. Group G on the other hand, used a column graph to illustrate their findings on whether grade 7 and 8 students thought condom machines should be installed in high school washrooms. However, a pie chart was also used to demonstrate that 73% of the students surveyed were male while only 27% were females. This graph was used by the group to indicate that their "data may have been a little wrong because more males were surveyed than females."

Generally, groups presented their projects without elaborating on the results, merely listing frequency values and/or percentages. Given the clear meaning of such data, elaboration may not have been necessary. However, explanations of the sampling procedure used were lacking in the presentations. Sampling issues such as sample size and representativeness were not discussed. However, when prompted by classmates about such information, groups were able to respond in a knowledgeable way. It was during these "question periods" that many of the students demonstrated knowledge that was not evident during their presentations. For instance, when asked by the teacher whether the results would differ had another sample been used, students in Group A responded that it depended on the level of the class: students in the same grade but attending more advanced classes would probably enjoy English and French classes more than the students sampled in their survey. Thus, although groups did not readily discuss sampling issues in their presentation, further prompting during the question periods revealed that this omission may have been due to oversight rather than to a lack of understanding. It was also during the question periods that confusion about particular concepts and graphs were demonstrated (e.g., Group C mentioned previously).

Structured Group Journals

Written protocols of group journals were analyzed in terms of the frequency and accuracy of groups' responses to prompts requiring that they: (a) define concepts; (b) explain the meaning of concepts; (c) reason about data and graphs; (d) plan for group projects; and (e) reflect on their learning. Given that there was a maximum number of prompts that could be responded to for each prompt type (i.e., category) and that this number varied across categories, the frequency data was converted to percentages for the purpose of analysis. As the response rates in Table 10 demonstrate, groups did not frequently make journal entries.

Given that all journal prompts except for planning were presented before groups were assigned to the two treatments, differences in response between the text and library of exemplars treatments were expected for the planning prompt only. A planned comparison was conducted to examine whether the library of exemplars tool was more effective in helping groups plan than the text approach. Results indicated that there were no significant differences between the library of exemplars and text treatments for planning (F(1, 22)= 1.602, p>0.05).

Table 10

Mean Percentage of Responses to Prompt Types in the Text and Library of Exemplars Treatments

		Treatments	
Prompt Type	Number of Prompts	Library of Exemplars (*n*=2)	Text (*n*=4)
Definition	11	0.319 (0.193)	0.364 (0.364)
Explanation	11	0.091 (0.000)	0.455 (0.234)
Reasoning	30	0.117 (0.023)	0.525 (0.050)
Planning	13	0.347 (0.054)	0.577 (0.263)
Reflection	12	0.083 (0.000)	0.125 (0.108)

Structured journals were also examined qualitatively to examine group learning throughout the tutorial sessions. Responses to prompts requiring that groups define concepts demonstrated that students acquired declarative knowledge of concepts such as population, sample, randomization, mean, mode, and median. Groups were also able to explain how to calculate the mean and range as well as explain why calculating the mean was important. The necessity of calculating the range, however, was not clear to students. The link between the mean and the range was not made by groups who explained that the range needed to be calculated in order to "find the difference between the numbers."

Despite the groups' ability to correctly define concepts, additional questions that required groups to reason about data based on an understanding of the mean revealed difficulties in distinguishing between the mean and the median. For example, when asked to explain why a particular value was obtained for the mean, some groups responded that it was due to its being the middle number, the average despite their previous definitions of the mean as the average and the median as the middle number. This suggested that perhaps the groups' difficulty in distinguishing between the median and mean may have been due to the choice of words used to explain these terms.

Prompts requiring that groups document plans for their group projects and those requiring they identify strengths and weaknesses were generally ignored. Ideas for group projects were documented, however, when prompted to document plans for computing analyses and creating graphs, groups were unresponsive.

Students' Evaluation of Instructional and Assessment Activities

Students evaluated the statistics instruction they received by responding to questions on the mini-course evaluation that examined: (a) the level of interest generated by the course; (b) the instructional and assessment activities which were identified as being the most effective for promoting statistical understanding; (c) the role of idea generation and revision; and (d) recommended changes to the course.

The mini-statistics course seemed to have stimulated much interest among students. According to overall ratings (1= liked it alot, 3= it was ok, 5= did not like it at all), the two-week mini-course was well liked by students (*M*=1.59, *SD*=0.87). Generally, the course was thought to be interesting and fun, "much more fun than mathematics." The most popular aspect of the mini-course was the use of computers. Sixty-five percent of students stated that they liked using computers for learning statistics and particularly for doing graphs. Students also enjoyed doing the project and working in groups. The least liked aspects of the course were the 10-15 minute lecture and homework assignments.

In addition to providing feedback regarding the level of interest generated by the course, the mini-evaluation was designed to encourage students to identify activities which were most effective in helping them understand statistics. Two types of activities and one kind of resource were identified by students as being invaluable for learning statistics: group presentations, activities providing hands-on learning experience, and mentoring. According to students, group presentations were effective in helping them gain a better understanding of statistics in that they: (a) depicted multiple ways of presenting data and graphs; (b) illustrated the data collection process; (c) aided in learning the meaning of terms such as mean, mode, and median; and (d) demonstrated the importance and value of research questions. Statistics activities which provided students with hands-on experiences that allowed them to learn statistical terms, manipulate and collect data, and create graphs were also deemed invaluable for comprehending statistics. The most important resource, however, was the experimenters who served as mentors during the tutorial and project planning.

The role of idea generation and revision was examined in responses to two questions about group projects. One question required that students state whether they kept their first project idea and if they changed it, to explain why they did so. Forty-seven percent of students stated that they did not keep their first idea because it was either too complex or not unique or interesting enough. The second question examined whether peer presentations influenced students' ideas for subsequent projects. Sixty-five percent of students stated that they would not change their group projects even after having watched other presentations. The remaining thirty-five percent of students wanted to change their projects for different reasons. For example, one student felt his group's project should have included a research question for making predictions, another would have made the project more creative, and a third student had new ideas that were stimulated from other presentations which would have been used to design a completely new study.

Of interest was whether or not there was consensus among individual students within a group about whether changes should be made to their projects. Students in three different groups unanimously agreed that they would not have changed their group projects even after having seen and critiqued their classmates' presentations. In contrast, students belonging to two other groups all agreed to change their projects for reasons mentioned above. However, no consensus was found among individual members of the remaining three groups. The majority of students in two of these non-consensus groups stated that they would have kept their project while the majority of students in the third group would have preferred to modify it.

Finally, recommendations for improving the mini-course varied but the most prominent was to make the course longer, perhaps spanning over a year. Other suggestions included more instruction to ensure understanding of statistical terms such as mode and median and explanations about projects earlier in the instruction.

CHAPTER 4: DISCUSSION

Five main questions were posed in the present research: (a) can statistics be taught at the eighth grade level, (b) can instruction and assessment of statistics be integrated to promote learning, (c) can small-group cooperation facilitate the learning of individual students, (d) does the extent to which assessment criteria are made explicit (i.e., text vs. library of exemplars approach) make a significant difference in learning, and (e) can authentic measures reliably assess learning. This chapter addresses these questions, identifies limitations of the study, and discusses educational implications and future research directions.

Can Statistics Be Taught At The Eighth Grade Level?

The present study demonstrates that a cognitive apprenticeship method of instruction facilitates statistical learning at the eighth grade level. Such instruction resulted in substantial knowledge gains from pretest to posttest. In addition, this method produced adequate performance on group presentations. The increase in knowledge is considerable given that statistics instruction is currently not in the secondary school curriculum and that the instruction in this study was only of a four-day duration. Overall performance on projects is also impressive despite the limited time (i.e., 3 days) students had to develop and conduct an experiment. Course evaluations indicated that length of instruction proved a concern. Although students "liked the mini-course alot," they felt it should be extended to a full year. Students' interest in statistics stemmed from the instruction's emphasis on interpretation rather than computation. By demonstrating that statistics involves more than number crunching, cognitive apprenticeships foster an awareness of and appreciation for the utility of statistics. In addition, course evaluations indicate that cognitive apprenticeships facilitate statistical understanding through mentoring.

In addition to providing information about students' overall knowledge of and interest in statistics, the present study examines students' understanding of (a) measures of central tendency, measures of variation, and anomalies and (b) population and sampling. This understanding is discussed in the following sections.

Measures of Central Tendency and Variation

Measures of central tendency. Research indicates that when formal methods of instruction are employed at the university level, students' understanding of the mean is limited to the computational formula: Conceptual understanding is not attained (Pollatsek et al., 1981; Zawojewski, 1988). The present study demonstrates that when a cognitive apprenticeship method of instruction is utilized at the eighth grade level, students' understanding of the mean, median, and mode is not restricted to formulas: However, nor is it conceptual. Students were able to define the mean and median in terms of "average" and "middle number" without using formulas. However, these definitions were used interchangeably suggesting that the concepts mean and median were not fully understood. Insufficient understanding was also demonstrated by students' inability to (a) calculate such measures on test essays despite knowing how to do so and (b) use appropriate measures for analyzing project data.

Students' inability to use formulas to calculate the mean, median and mode by hand can be attributed to the instruction's emphasis on interpretation rather than computation. This emphasis may account for the different results found by Pollatsek et al. (1981). Contrary to the instruction in Pollatsek et al.'s (1981) study, students in this study were never given formulas to memorize. Rather, computer software was used as a tool to solve computational problems in order to support conceptual understanding (Lajoie, 1993; Salomon et al., 1991). This tool, however, was not provided on test essays. Students were consequently unable to transfer their knowledge of how to calculate the mean on problems that required they perform such computations by hand. This finding suggests that students did not fully grasp how to use the procedures. As Resnick (1989) argues, students require sufficient understanding of content in order to become effective problem solvers. This study demonstrates that adequate understanding of measures of central tendency requires an ablility to compute and interpret data. Perhaps providing opportunities to discuss and compare results obtained through hand calculations with those procured using data analysis software can foster such understanding.

Sufficient understanding of measures of central tendency also requires an ability to select measures based on the nature of the data to be analyzed. This understanding was not demonstrated in the present study. One group used all measures (i.e., mean, median, and mode) regardless of whether these were apt for analyzing frequency data collected on projects. Most groups used percentages to analyze this type of data, however, it is unclear whether this choice was based on an understanding that measures of central tendency were inappropriate or on an interest in constructing pie graphs which automatically generated these percentages. Insights are provided by Lavigne, Lajoie, Munsie, and Wilkie (1994) whose case study of one of these group's discussions during project design revealed that choice of measures was determined by what was most appealing (e.g., doing graphs) rather than what was appropriate. Although Lavigne et al.'s (1994) findings are not generalizable, they do highlight the importance of examining interactions during planning. Such an examination in the present study would have shed light on the selection criteria groups used for their choice of analysis.

The present study highlights the need to make connections between data computation and interpretation. The need to instruct students about the relationship between data organization, data analysis, and graphic representation is similarly accented. Informal observations indicated that in many cases, students did not know how to organize data in a format appropriate for constructing their pie graphs. Students were not attuned to the types of variables they were working with (e.g., gender) and consequently entered data by trial and error. Such problems are atributed to insufficient instruction on how to organize data for different problems. This observation supports Hancock, Kaput, and Goldsmith's (1992) finding that statistical understanding requires explicit instruction of data organization and variable characteristics.

Measures of variation. One measure of variation was taught in the present study: Range. Given tha: the instruction focused on interpretation rather than computation, one would expect students to acquire a conceptual understanding of range. This expectation was not met. Students did acquire significant knowledge of range, however, this knowledge was limited to the computational formula. As demonstrated by journal entries, the purpose of range was "to find the difference between numbers" rather than to provide a variability index which reflects the accuracy of a measure such as the mean. This finding is consistent with Pollatsek et al. (1981) who found that students comprehend the mean--in this case the range--in terms of an abstract formula that has no meaning. It is unclear why students in the present study were unable to extend their knowledge beyond the computational formula.

Anomalies. Instruction of anomalies in the current study focused on the effect of outliers on the mean. Students' understanding of the concept, however, was limited despite instruction. Most students were unable to define or illustrate an outlier using a chart or graph. In addition, the effect of outliers was not discussed in any of the presentations, including by the group that used the mean to analyze their data. Since survey data was collected and small samples were used, it may be that students did not discuss outliers in their presentation because none were identified. Informal observations during group discussions suggest that groups did not examine their data for outliers.

Summary. The present study demonstrates that students failed to acquire a conceptual understanding of the mean, median, mode, range, and outlier despite instruction. Explanations of measures of central tendency lacked depth and students' understanding of the range was limited to the computational formula. One suggestion for fostering procedural knowledge and conceptual understanding of the mean, median, and mode is to provide opportunities for computation and interpretation by allowing students to do manual computations and use data analysis software. Providing opportunities for data organization and the selection of appropriate measures is also deemed important to promote statistical understanding.

Population and Sampling

This section discusses students' understanding of population and sampling in terms of (a) population and sample, (b) randomization, and (c) sample size and representativeness.

Population and sample. Students had a working knowledge of population and sample which was demonstrated in everyday definitions such as "a population consists of a large group of people" and "a sample is an example or subset of something." However, students' understanding of population did not go beyond this everyday definition despite instruction. Students were unable to identify a population from problem scenarios. The notion of population was not referred to in presentations or on essays that required an explanation of how to collect data given a particular research question. Students' understanding of sample, however, was significant as a result of instruction. Samples were identified from problem scenarios. In addition, the term sample was incorporated into students' vocabulary and used readily to explain data collection. However, despite this understanding students were unable to make inferences from a sample to a population. Although the relationship between sample and population was explained in the instruction, inferences based on this relationship were not explicitly modeled. This neglect may also account for the lack of increase in students' understanding of population despite instruction. Providing instruction for learning the part-whole relationship between sample and population may improve students' understanding of sampling. According to Schwartz et al. (1994), such instruction can help students encode situations in statistical terms.

Randomization. According to Konold, Lohmeier, Pollatsek, Falk, and Lipson (1991), the notion of randomness is at the heart of probabilistic and statistical reasoning. Randomization devices are often utilized to teach or examine the learning of randomness. In the present study, however, randomization was modeled by having students participate in an activity that involved controlling for bias through random assignment. When immediately asked to explain why such a procedure was important, students stated that it ensured "fairness." However, by the end of the instruction, students were unable to apply this knowledge to (a) explain randomization in their own words, (b) provide a conceptual (e.g., to ensure fairness) rather than procedural (e.g., to pick at random) definition, and (c) consider and apply randomization techniques when designing and conducting a mini-experiment. Although random selection was also discussed in the classroom, students failed to randomly select subjects from the pool of eighth grade students. Rather, data was collected from an immediate pool of subjects; i.e., classmates. Comments by students suggest that sampling may have been driven by practical considerations, such as time allotted to plan and conduct a mini-experiment. Students may have simply resorted to sampling their classmates because it was easier and faster.

Sample size and representativeness. The present study demonstrates that eighth grade students can acquire a good understanding of sampling in terms of the size and representativeness of a sample. For instance, larger samples were understood to provide more precise estimates than smaller samples and characteristics of a sample such as gender, were known to influence results. However, despite this knowledge, students failed to apply such principles when collecting data for their experiment. Practical issues overshadowed considerations of size and representativeness of a sample. For instance, when questioned about the size of their samples during presentations, students acknowledged that their samples were small but explained that collecting more data would have been too time consuming. Similarly, one group recognized the bias in their sampling of male and female subjects but explained that this was due to insufficient time in which to sample more males.

The finding that students' sampling decisions were greatly influenced by practical considerations emerged as a result of "question periods" which followed group presentations. Question periods provided classmates with an opportunity to ask for clarification, elaboration, and justification (Lampert, 1990). It was during these sessions that students' understanding of sampling issues such as size and representativeness of a sample were made more explicit. Most groups, except for the one mentioned above, did not readily discuss the issue of sample representativeness in oral presentations. However, requests for clarification and elaboration during question periods revealed that this was due to omission rather than to a lack of understanding. For instance, when queried about the use of different samples, one group explained that if students streamed to higher-ability courses had been sampled, their results of favorite subject matters would have differed substantially. Another group explained why their results would have differed had they sampled adults or students from grades 9 to 11 rather than from grades 7 and 8. These findings demonstrate that in addition to communicating knowledge, students must be able to clarify and elaborate upon request. Such requests can serve as concrete prompts that scaffold learning (Rosenshine & Meister, 1992). Written and verbal prompts can be used in such a way. However, findings from the present study suggest that written prompts are not as effective as verbal prompts. Groups' unresponsiveness to written prompts in the structured journal leaves doubt as to whether omissions were due to a lack of understanding or to insufficient documentation. More adequate ways of prompting learners to demonstrate their knowledge on journals may be required. This suggestion reflects Resnick's (1989) contention that question posing must be carefully engineered to promote productive discourse in cooperative settings. On the other hand, responses to verbal prompts such as those presented in the form of requests for clarification and elaboration, reveal an understanding of sampling issues that were not explicitly demonstrated in presentations. Direct measures such as verbal protocols are therefore more effective for assessing learning than are indirect measures such as journals. This finding is consistent with Jacobs (1993) who found that the richness of students' reasoning skills, although masked on paper and pencil tests, was revealed during verbal interactions. In the case of the present study, it may be that verbal prompts which require students to do more than articulate their knowledge encourage students to think critically and to articulate their knowledge more clearly.

Investigating discussions arising from pertinent questioning can highlight factors that influence learning. In the current study, prompting revealed the practical considerations motivating students' decisions about sampling. Other researchers have found that opinions serve as a barrier to adult and middle school learners' understanding of statistical principles such as representative sampling (Jacobs, 1993; Schwartz et al., 1994; Tversky & Kahneman, 1971). The adverse effect of opinions on statistical understanding was not found in the present study. This can be attributed to the fact that verbal protocols during problem solving activities were not examined. Such an examination in the present study would have shed some light on the role of everyday knowledge on statistical learning. Everyday knowledge enabled students to solve problems in which the solution was consistent with their real-life experiences. However, verbal and written responses to problems where the solution is inconsistent with students' everyday knowledge would provide more detailed information about the role of everyday knowledge on students' statistical understanding. As Fong et al. (1986) and Jacobs (1993) contend, such problems can challenge students' to confront their intuitive notions and assumptions with statistical evidence.

Summary

As this study demonstrates, cognitive apprenticeship can be used to teach statistics to eighth grade students in a way that promotes learning. Although students understanding of measures of central tendency was limited, their understanding of sampling was strong. This is encouraging given the abstract nature of such concepts. However, the generality of their responses suggest that depth of understanding had not been acquired. Given the content coverage and the time in which students were required to learn concepts, it is not surprising that breadth rather than depth of knowledge was acquired. Yet, learning in contemporary statistics classes is characterized by such instruction. University students lack sufficient background to fully master statistical content (Posten, 1981), yet the coverage in statistics courses is high in relation to the time and number of courses available to learn content. Providing statistics instruction earlier in the curriculum as recommended by NCTM (1989) could circumvent the content coverage problem and allow for in-depth exploration of concepts over a longer time period. Projects have potential for such exploration. However, allowing three days to complete a project consisting of designing and conducting a mini-experiment may be unrealistic. Practical considerations can limit the extent to which statistical procedures can be applied on such a task. Opportunties for articulating knowledge as well as clarifying, elaborating, and justifing ideas during group interactions can highlight other factors which may affect or contribute to the learning of statistics.

Can Instruction and Assessment of Statistics Be Integrated To Promote Learning?

Instructional activities and assessment tasks in the present study were designed to be consistent with recommendations made by NCTM (1989) and ASA (1991). In this sense, instruction and assessment were integrated. Further integration can be obtained in dynamic assessment where immediate feedback serves as a form of instruction (Lajoie & Lesgold, 1992). Such integration was reflected in one assessment task of this study: Oral presentations of group projects. The notion was that sharing ideas with peers through oral presentations would provide learners with an opportunity to revise their thinking and engage in self-assessment (Lampert, 1990). The findings of the present research support this notion to some degree. For example, the first group to present their project expressed a desire to modify and redo their presentation once all other groups had presented. This was surprising given that the group had rated their presentation a perfect score of 50 and stated that this was an accurate reflection of their performance. When queried by the teacher regarding their reasons for requesting a second opportunity to present their project, the group responded that they felt they "had not done a very good job presenting." Comparisons with other groups may have motivated this group's desire to redo their presentation. Such comparisons were also evident when other groups stated that they wanted to present last because they felt that their presentations were not yet at the level of other groups' presentations.

As demonstrated above, some groups re-evaluated their performance as a result of watching presentations given by classmates. This self-assessment, however, referred to the presentations themselves rather than to the projects. Project presentations were identified by students as one of the most beneficial activities for helping them understand statistics. However, although motivated to change their presentations, comments on course evaluations indicate that students would not have changed the actual projects. These comments suggest that although group projects were beneficial for learning statistics, the new knowledge acquired during these presentations did not necessarily motivate students to revise their projects. Self-assessment seemed geared towards the clarity of the presentation rather than to the quality of the project. Perhaps this finding is due to the fact that group projects were similar in every respect except for the research question examined. Projects based on similar ideas would be less likely to promote revisions than projects which demonstrate a different perspective.

Projects were considered more beneficial for learning statistics than were paper and pencil tests such as homework assignments. This comment reflects the dynamic nature of project assessments and is consistent with Diez and Moon's (1992) view that assessment criteria are guides to learning rather than the terminal point to learning. It is also possible that shared responsibilities incurred by group collaboration in projects reduced anxieties associated with performing. The fact that assignments had to be completed individually also may have contributed to students' dislike of the task.

In conclusion, it seems that projects and oral presentations have much potential for examining thinking skills such as problem solving, reasoning, communication, and connectedness. Although this part of the study focused on communication, the findings suggest that project-based learning is large in scope, allowing for the use and integration of a variety of knowledge and skills.

Can Small-Group Cooperation Facilitate the Learning of Individual Students?

Although the present study examined both individual and group performance to obtain a comprehensive and valid picture of learning (NCTM, 1989, 1993; Webb, 1993), the data allowed for an informal rather than systematic examination of whether small-group cooperation facilitated the learning of individual students. Individuals' test scores suggest that cooperation in mixed-ability groups may be beneficial for all ability levels. Although knowledge gains of students within each group were relatively similar, group performance on oral presentations of projects suggest that some students were more knowledgeable than others. These students were the ones to respond to inquiries and elaborate on results. These findings are consistent with Webb (1993) who found that both group and individual measures are required for valid assessments of learning. Although Webb found that students' learning was overestimated in group assessment, the present study suggests that such problems can be minimized through oral presentations which require students in each group to discuss various aspects of their project. Such discussions can enable teachers to identify the nature of student's understanding within the group.

Information about learning in cooperative settings is incomplete unless the nature of cooperation is examined. The present study demonstrates that interpersonal conflicts between group memebers can negatively affect learning. Grouping based on friendship, however, does not necessarily result in effective learning. Some "friends" in the study engaged in off-task behavior when working together which limited their learning. In addition, contrary to Webb (1991) who found that high ability students were more likely to help others, the present study suggests that such students do not necessarily have the patience to provide assistance. One high ability student insulted rather than helped peers who were slower in understanding some of the concepts and procedures. Such behavior discouraged lower-ability students from participating in activities and impeded their learning of statistics. Since the effect of cooperation in the present study is examined informally and based on observations of two groups, the findings are merely suggestive and thus not generalizable. However, the observations support Cohen (1994) and Webb's (1991) recommendation that successful learning in cooperative environments require that students be trained to work cooperatively, give explanations to each other, and be sensitive to students' need for help.

Does The Explicitness of Assessment Criteria Make A Significant Difference In Learning?

The present study provided transparent criteria for assessing project work to ensure that learners adapted their performance accordingly (Frederkisen & Collins, 1989) and that expert judgements of performance were made (Diez & Moon, 1992). According to Frederiksen & Collins (1989), making assessment criteria transparent enables learners to assess themselves and others with almost the same reliability as evaluators. Transparency in the present study was provided in two treatment tools: the library of exemplars and text stacks. Both tools provided explicit standards for performing and assessing group projects by (a) specifying and defining criteria for designing and conducting an experiment based on NCTM (1989) standards and ASA (1991) guidelines and (b) identifying the value of performance on each criterion. However, the library of exemplars provided groups with an added dimension of explicitness not provided in the text stack: benchmarks for success. Such benchmarks modeled performance and elaborated on the assessment criteria through examples of average and above average performance. Given that yardsticks of performance were provided in the library of exemplars but not in the text stack, it was hypothesized that students in the former treatment would (a) be more likely to plan projects, (b) assess performance in the same way as experts, (c) receive higher project ratings, and (d) receive higher test scores.

The hypothesis that students receiving the library of exemplars treatment would be more likely to plan projects than students receiving the text treatment was not confirmed. Treatment differences were not found in the mean percentage of responses made to planning prompts in structured journals. Response rates were low for both treatments. The limited time delay between the presentation of the two tools and the commencement of projects may have affected groups' responsiveness to planning prompts. In addition, the limited time in which to complete projects may have deterred groups from doing any planning. Perhaps the stacks should have been made available earlier in the instruction to allow for sufficient planning prior to the commencement of projects.

The hypothesis that the library of exemplars approach would provide more similar ratings between students and experts (i.e., experimenters) than the text approach was not confirmed. No significant treatment differences were found between ratings of experimenters and groups or between ratings of experimenters and self. However, overall mean ratings for self, group, and experimenters indicate that students' self-assessments were quite similar to experimenter ratings but that group ratings were lower than both self and experimenter assessments. These findings suggest that making criteria transparent through examples or text is sufficient for facilitating a closer alignment of students' self-assessment with evaluators' assessment. The fact that overall group ratings were lower than other types of ratings suggests that group competition may have played a role in how reliably students assessed others. Competition seems likely given previous comparisons which resulted in students' desire to either redo their presentation or to present last.

The hypothesis that students receiving the library of exemplars treatment would receive higher project ratings than learners in the text treatment was not confirmed. Groups in the library of exemplars and text treatments performed equally well on presentations. Specifying assessment criteria and providing explanations of what is meant by such criteria seems sufficient for learners to adapt their performance to standards set by instructors. Finally, the hypothesis that students receiving the library of exemplars treatment would outperform those in the text approach on test essays was not confirmed. Students in both treatments performed well as demonstrated by significant knowledge gains. Such significance suggests that the presence of any form of elaboration of assessment criteria (i.e., text or examples) may be sufficient for facilitating learning. However, post-hoc analyses of test performance on items addressing each concept demonstrates that elaboration of criteria through exemplars is more effective for faciliating understanding of sample representativeness than through text. This effectiveness was also found in group presentations where representative sampling was discussed by a group assigned to the library of exemplars. These findings suggest that learning can be facilitated by providing multiple representations of realistic performance to make abstract concepts such as representative sampling more concrete. In addition, although limited by a small sample size, results from the second homework assignment suggest that providing examples (i.e., library of exemplars treatment) of performance where concepts and procedures are integrated may be more effective for promoting connectedness than text (i.e., text treatment).

Although the library of exemplars tool was effective at promoting some learning, the reduced effect of the library of exemplars approach may be attributed to two factors: (a) low volume level of the sound from digitized video clips in the library of exemplars and high noise level in the class and (b) insufficient differentiation between exemplars. The volume level of the library of exemplars was designed for a moderately noisy environment. However, the noise level in the classroom increased substantially as groups discussed potential projects and worked on computers. Students quickly lost interest as they had to repeatedly replay clips in order to hear the muffled sound emanating from the video clips. Such noise may have contributed to the unexpected results. In addition, exemplars in the library may not have been rich enough to clearly differentiate between the two levels of performance. Perhaps the significant effect of the library of exemplars for representative sampling was due to richer exemplars whereas the marginal effect of the library for other concepts may have been due to insufficient differentiation between performance levels. Buiding a richer library of exemplars might provide some insights about the role of such exemplars on learning.

Can Authentic Measures Reliably Assess Learning?

The reliability of authentic measures is of particular concern given that they lack sufficient psychometric foundations. The subjectivity of scoring performance on authentic measures (Hacker & Hathaway, 1991) can affect the reliability of assessment unless (a) templates are developed to objectively assess the quality and accuracy of performance and (b) multiple raters are used to score performance (Brandt, 1992). In the present study, high inter-rater reliabilities for pretest (r=0.982), posttest (r=0.987), assignment one (r=0.874), and assignment two (r=0.946) indicate that authentic measures such as open-ended paper and pencil tests can be used to reliably assess performance when two judges are used. Achieving high inter-rater reliability for assessing performance on ill-defined tasks such as projects, however, is more difficult. Although assessments in this study were made clear to evaluators (i.e., experimenters and groups of students) and performance was assessed by multiple judges (4 and 7 respectively), separate inter-rater reliabilities on experimenters' and groups of students' overall ratings ranged from low to moderately low. These results indicate that project assessments were generally inconsistent regardless of whether the evaluator was an experimenter or learner.

Given that the assessment criteria were clearly and explicitly defined to both types of evaluators, the low inter-rater reliabilities may be due to an unclear correspondence between the type of scoring used to train raters and the type of rating used to assess performance. The training received by raters was based on average and above average examples of performance on each criterion yet scoring was based on the allocation of points (i.e., 1-5 or 1-10). This ambiguity may have resulted in the low reliability between experimenters and between groups of learners. The difference between a score of 5 and a score of 7 on a criterion with a maximum rating of 10, for example, is considerable yet both scores fall in the category of average performance. These findings suggest that inter-rater reliability requires more than clear assessments and multiple raters. Benchmarks of performance on each criterion should be specified according to the type of scoring that will be required of evaluators. This correspondence was clearer in templates developed for scoring responses on authentic paper and pencil measures (i.e., tests and assignments). These templates were consequently more effective in attaining high reliability than were guidelines for assessing projects.

Overall inter-rater reliabilities, however, are not completely informative when they are moderately low or moderately high. It may be that some criteria are more difficult to score than others. As demonstrated in this study, experimenters had difficulty scoring performance on half of the criteria. These criteria included data collection, data analysis, and presentation style. Given the complex nature of the first two procedures and the ambiguity of the last criterion, this finding is not surprising. Sampling and data analysis procedures must therefore be clearly defined to reliably assess performance. Multiple examples of different aspects of such procedures may also help anchor their meaning and make assessments less ambiguous and more reliable. As demonstrated in this study, performance that explicitly conforms to the criteria are also easier to score. Experimenters agreed more when scoring the presentations of groups in the library of exemplars treatment than the presentations of groups in text treatment which indicates that the former presentations were clearer and thus more easier to score than the latter.

In conclusion, this study demonstrates that authentic measures can reliably assess performance, particularly when templates are developed to assess open-ended paper and pencil tests. In this case, only two raters were sufficient for attaining high inter-rater reliabilities. Reliably assessing performance on more ill-defined tasks such as projects, however, is more difficult. The scoring system must be consistent with all levels of performance on the criteria in question. Moreover, criteria that refer to abstract concepts and complex procedures must be clearly defined. Given the difficulties in assessing performance on projects, at least four raters are required to ensure acceptable inter-rater reliabilities.

Limitations of the Study

One of the major limitations of the present study is the noise and chaos associated with conducting research in naturalistic settings. The noise level within the classroom reduced the full impact of the library of exemplars tool by muffling the sound emanating from the digitized video clips. The benefits of the library of exemplars approach are consequently unclear. A second limitation is the duration of the instruction which was insufficient for providing in-depth instruction of statistical content. A third limitation of the study is the ineffectiveness of tutorial prompts to promote discussions. Although students were constantly reminded to respond to these prompts as a group, many simply ignored them. Such omissions had two consequences: (a) impeded the development of students' reasoning skills and (b) limited conclusions about learning which were based on information from structured journals and audiotapes. A fourth limitation is the ineffectiveness of homework assignments and the structured journal in assessing learning. Low response rates severely limited the investigation of on-going learning. This problem was compounded by the fact that verbal protocols which generally serve as on-line measures of learning were not examined. In addition, the full impact of project-based learning environments was limited by the amount of time given to students for completing a project. A fifth limitation is insufficient collaboration between the teacher and researcher in the development of assessment measures which resulted in ambiguities in the wording of problems on assignments (e.g., assignment two) and essays. Finally, this study is limited by the insufficient alignment of performance descriptors with the type of rating expected of evaluators which substantially reduced inter-rater reliabilities.

Educational Implications and Future Directions

The present research has four implications for teaching statistics and assessing learning. First, cognitive apprenticeships can be used to promote motivation and learning. However the nature of the knowledge acquired depends on the amount of content covered in the instruction. Overwhelming students with too many concepts within a limited time frame results in the acquisition of breadth rather than depth of knowledge. Perhaps longitudinal research can determine the content coverage appropriate for promoting depth of understanding. A first step might be to conduct research on the learning of statistics spanning over a year. Second, engaging students in articulation, clarification, elaboration, and justification through verbal discussions in oral presentations is critical for (a) obtaining valid information about student learning and (b) engaging students in self-assessment. Teachers must therefore be committed to (a) probing students' knowledge in a variety of ways and (b) developing multiple assessments that measure learning directly. However, researchers examining the effectiveness of alternative methods for probing knowledge must collaborate with teachers if these methods are to be employed in classrooms. Third, learners can more easily learn abstract concepts such as sample representativeness when visual exemplars that demonstrate how to perform representative sampling concretely are presented. Richer examples of such performance might be used to further develop the library of exemplars. Finally, cooperative learning in heterogeneous groups can be effective if students are given training prior to instruction on how to work cooperatively and on how to provide assistance to peers. Providing such training can minimize the influence of personality factors on the productivity and effectiveness of cooperative groups. Future research might include a case study which controls for all the noise encountered in the present study. A small group of students learning statistics as part of the curriculum could allow for extensive and intensive examination of learning on measures which might include (a) evaluating and critiquing statistical results presented in the media, (b) participating in debate sessions, and (c) teaching an instructional lesson on a particular concept. Such research would provide valuable information about the nature of statistical learning on a variety of direct measures.

References

Adobe Systems. (1991). *Adobe Premier Version 1.0* [Computer program]. Mountain View, CA: Author.

Adobe Systems. (1992). *Adobe Photoshop Version 2.5.1* [Computer program]. Mountain View, CA: Author.

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

Apple Computer Inc. (1989). *QuickTime Version 1.5* [Computer program]. Cupertino: CA: Author.

Archibald, D. A., & Newman, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in secondary schools.* Washington, DC: National Association of Secondary School Principals.

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational Psychologist, 26* (3 &4), 369-398.

Brandt, R. (1992). On performance assessment: A conversation with Grant Wiggins. *Educational Leadership, 49* (8), 35-37.

Brandt, R. (1989). On misuse of testing: A conversation with George Madaus. *Educational Leadership, 46* (7), 26-29.

Claris Corporation. (1991). *Hypercard Version 2.1* [Computer program]. Santa Clara, CA: Apple Computer Inc.

Claris Corporation. (1987). *MacPaint Version 2.0* [Computer program]. Santa Clara, CA: Apple Computer Inc.

Cognition and Technology Group at Vanderbilt. (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher, 19* (5), 2-10.

Cognition and Technology Group at Vanderbilt. (1992). The Jasper series as an example of anchored instruction: Theory, program description, and assessment data. *Educational Psychologist, 27* (3), 291-315.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64* (1), 1-35.

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.* Hillsdale, NJ: Earlbaum.

Collins, A., Hawkins, J., & Frederiksen, J. R. (1991). *Three different views of students: The role of technology in assessing student performance* (Report No. 12). New York: Center for Technology in Education.

Costa, A. L. (1989). Re-assessing assessment. *Educational Leadcrship, 46* (7), 1.

Cricket Software Inc. (1989). *Cricket Graph Version 1.31* [Computer program]. Valley Stream Parkway, PA: Author.

de Lange Jzn, J. (1991). Assessment: No change without problems. In T. Romberg (Ed.), *Reform in school mathematics and authentic assessment.* NY: SUNY Press.

de Lange Jzn, J., & Verhage, H. (1992). *Data Visualization.* Scotts Valley, CA: Wings for Learning.

Diez, M. E., & Moon, C. J. (1992). What do we want students to know?... and other important questions. *Educational Leadership, 49* (8), 38-41.

Duren, P. E., & Cherrington, A. (1992). The effects of cooperative group work versus independent practice on the learning of some problem solving strategies. *School Science and Mathematics, 92* (2), 80-83.

Farallon Computing Inc. (1990). *MediaTracks: Version 1.0* [Computer Program]. Emeryville, CA: Author.

Farallon Computing Inc. (1990). *ScreenRecorder: Version 2.0* [Computer Program]. Emeryville, CA: Author.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics, 15*, 1-24.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253-292.

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

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19* (1), 44-63.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist, 44* (2), 134-141.

Hacker, J., & Hathaway, W. (1991, April). *Toward extended assessment: The big picture*. Paper presented at the annual meeting of the Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Hamm, M. (1992). Achieving scientific literacy through a curriculum connected with mathematics and technology. *School Science and Mathematics, 92* (1), 6-9.

Hancock, C., Kaput, J.J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27* (3), 337-364.

Herman, J. L. (1992). What research tells us about good assessment. *Educational Leadership, 49* (8), 74-78.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80* (4), 237-251.

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

Kapadia, R. (1982). A practical approach to statistics. In Organising Committee of the First International Conference on Teaching Statistis (Ed.), *Proceedings of the First International Conference on Teaching Statistics, 1*, (pp. 169-178). Sheffield, England: Organising Committee of the First International Conference on Teaching Statistics.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Monterey, CA: Brooks/Cole Publishing Company.

Kirst, M. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Leadership, 20* (2), 21-23, 27.

Konold, C. Lohmeier, J., Pollatsek, A., Well, A., Falk, R., & Lipson, A. (1991). Novice views on randomness. *Proceedings of the Thirteenth Annual Meeting of the International Group for the Psychology of Mathematics Education - North American Chapter, 1*, 167-173.

Lajoie, S. P. (1993). Computer environments as cognitive tools for enhancing learning. In S. P. Lajoie & S. Derry (Eds.), *Computers as cognitive tools* (pp. 261-288). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lajoie, S. P. (1991). A framework for authentic assessment in mathematics. *National Center for Research in Mathematical Sciences Education, 1* (1), 6-12.

Lajoie, S. P., & Lesgold, A. (1992). Dynamic assessment of proficiency for solving procedural knowledge tasks. *Educational Psychologist, 27* (3), 365-384.

Lajoie, S. P., Lawless, J., Lavigne, N. C., & Munsie, S. D. (1993, April). *New ways to measure skills of problem solving, reasoning, communication, and connectedness.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Lampert, M. L. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal, 27* (1), 29-63.

Lavigne, N. C., Lajoie, S. P., Munsie, S. D., & Wilkie, T. V. (1994, April). *Authentic assessment of statistical reasoning in cooperative learning environments.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Linn, R. L. (Ed.). (1989). Current perspectives and future directions (pp. 1-10). *Educational Measurement, 3rd edition.* New York: Macmillian Publishing Company.

Linn, R. L., Baker, E L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Meyer, C. A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership, 49* (8), 39-40.

Moody, D. (1991). Strategies for statewide student assessment. *Policy Briefs, 17,* 1-5.

Mosteller, F. (1980). Classroom and platform performance. *The American Statistician, 34* (1), 11-17.

Mosteller, F. (1988). Broadening the scope of statistics and statistical education. *The American Statistician, 42* (2), 93-99.

National Council of Teachers of Mathematics Commission on Standards for School Mathematics (1993). *Assessment standards for school mathematics: Working draft.* Reston, VA: Author.

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

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90* (4), 339-363.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational Measurement, 3rd edition* (pp. 447-474). New York: Macmillian Publishing Company.

Pereira-Mendoza, L., & Swift, J. (1981). Why teach statistics and probability - A rationale. In A. P Shulte, & J. R. Smart (Eds.), *Teaching Statistics and Probability Yearbook* (pp. 1-7). Reston VA: National Council of Teachers of Mathematics.

Phelps, E., & Damon, W. (1989). Problem solving with equals: Peer collaboration as a context for learning mathematics and spatial concepts. *Journal of Educational Psychology, 81*(4), 639-646.

Pollatsek, A., Lima, S., & Well, A. (1981). Concept or computation: Students' misconceptions of the mean. *Educational Studies in Mathematics, 12,* 191-204.

Posten, H. O. (1981). Review of statistical teaching materials for 11-16-year olds. *The American Statistician, 35* (4), 258-259.

Resnick, L. B. (1989). Treating mathematics as an ill-structured discipline. In C. I. Randall & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 32-60). Reston, VA: National Council of Teachers of Mathematics.

Resnick, L. B. (1987). Learning in school and out. *Educational Researcher, 16,* 13-20.

Rogoff, B. (1991). Social interaction as apprenticeship in thinking: Guidance and participation in spatial planning. In L. B. Resnick, J. M. Levin, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 349-364). Washington, D. C.: American Psychological Association.

Romberg, T. A., Zarinnia, E. A., & Collis, K. E. (1990). A new world view of assessment in mathematics. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 21-38). Washington, D. C.: American Association for the Advancement of Science.

Rosenshine, B., & Meister, C. (1992). The use of scaffolds for teaching higher-level cognitive strategies. *Educational Leadership, 49* (8), 26-33.

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20* (3), 2-9.

Schwartz, D., Goldman, S., Moore, A., Zech, L., Smart, L., Mayfield-Stewart, C., Vye, N., & Barron, L. (1994, April). *Adolescent understanding of sampling in the context of a survey.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shaughnessy, J. M. (1982). Misconceptions of probability, systematic and otherwise; teaching probability and statistics so as to overcome some misconceptions. In Organising Committee of the First International Conference on Teaching Statistis (Ed.), *Proceedings of the First International Conference on Teaching Statistics, 2,* (pp. 784-801). Sheffield, England: Organising Committee of the First International Conference on Teaching Statistics.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook for research in mathematics teaching and learning* (pp. 465-494). New York: Macmillan Publishing.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments - Political rhetoric and measurement reality. *Educational Researcher, 21* (4), 22-27.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46* (7), 4-9.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20* (7), 2-16.

Singer, J. D., & Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statastician, 44* (3) , 223-230.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6* (3), 33-42.

Supermac Technology. (1991). *ScreenPlay Version 1.1.1* [Computer program]. SunyVale, CA: Author.

Systat Inc. (1988). *Mystat Version 1.0: A personal version of Systat.* [Computer program]. Evanston, IL: Author.

Tanner, M. A. (1985). The use of investigations in the introductory statistics course. *The American Statistician, 39* (4), 306-310.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76* (2), 105-110.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* , *5* , 207-232.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review, 90* (4), 293-315.

Varga, T. (1982). Statistics in the curriculum for everybody: How young children and how their teachers react. In Organising Committee of the First International Conference on Teaching Statistics (Ed.), *Proceedings of the First International Conference on Teaching Statistics, 1*, (pp. 71-80). Sheffield, England: Organising Committee of the First International Conference on Teaching Statistics.

Vykotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Watts, D. G. (1991). Why is introductory statistics difficult to learn? And what can we do to make it easier? *The American Statistician, 45* (4), 290-291.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics, 22* (5), 366-389.

Webb, N. (1993). *Collaborative group versus individual assessment in mathematics: processes and outcomes.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Whitehead, A. N. (1929). *The aims of education.* New York: MacMillan.

Wiggins, G. (1990). *The case for authentic assessment.* (Contract No. R-88-062003). Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproductive Service No. ED 328 611)

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49* (8), 26-33.

Williams, S. M. (1992). Putting case-based instruction into context: Examples from legal to medical education. *The Journal of the Learning Sciences, 2* (4), 367-427.

Wolf, D., Bixby, J., Glenn III, J., Gardener, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

Zawojewski, J. S. (1988). Teaching statistics: Mean, median, and mode. *Arithmetic Teacher, 35* (7), 25-26.

Appendix A

Summary of Tutorial Activities and Reasoning Prompts

An Introduction to Statistics

Activity: **Group Pulse Rates: Introduction to Data, Mean, and Range.**

The concept data was introduced to students in the context of taking pulse rates. Two levels of the variable pulse rates were examined: pulse rates at rest and pulse rates after a physical activity (i.e., running on the spot for one minute). Students were required to compute the mean and range for both levels and to compare the results.

Prompts:

1. What do the results mean?
2. How do pulse rates "At Rest" differ from "Runners"?
3. Discuss this with your group

Measures of Central Tendency

Activity: **School Grades: Mean and Extreme Scores**

Students were required to enter the following data in the computer: 50, 52, 56, 94, 98, and 100. The objective of this activity was to have students discuss the influence of extreme scores on the mean. Thus, before analyses were performed students were asked to eyeball the data and to describe what they saw. Once the mean was computed, students were prompted to interpret the results.

Prompts:

1. What information does this data provide?
2. Why do you think we obtained this mean?
3. Do you think this value reflects everyone's grade accurately?

4. Did anyone get 75% as a grade?

Activity: **School Grades: Mean and Outliers**

Students were given the following data to enter into the computer: 12, 74, 78, 89, and 92. This activity builds on the previous one by introducing other types of data that influence the mean: outliers. The goal again, was to emphasize the importance of examining data prior to analyses. As such, students were asked to predict the value of the mean for this particular dataset.

Prompts:

1. What do you think the mean will be for this dataset?
2. Do you think the value reflects everyone's grade?
3. Was the mean what you predicted?
4. What does this value tell you about your data?
5. How could the mean be more reflective of the data?
6. Which one of the grades in the dataset is called the outlier?
7. Why do you think it's called that?

Activity: **Group Pulse Rates: Introduction to Mode and Median**

Using the same pulse rate data that was collected in an earlier activity, students were asked to sort the data and to find the mode and median. The purpose was to make students aware that their choice of statistics depends on the type of data that they have and that eyeballing the data is therefore important.

Prompts:

1. Does any number occur more than once?
2. What is the median for this data?

3. Why do you think it might be important to get an accurate glimpse of what a group of numbers (data) might mean?

Measures of Dispersion

Activity: **School Grades: Range**

This activity builds on the previous two school grades activities which examined the influences of extreme scores and outliers on the mean to introduce the concept "range." The same datasets are used as the basis for discussion in this activity.

Prompts:

1. What is the range of scores for the data?
2. Do you think it's better to have a small range or a large range of scores, and why?

Sampling

Activity: **Class Pulse Rates: Introduction to Sampling**

Students are required to open their pulse rate datafile and enter the "at rest" pulse rates of the entire class, thereby increasing the sample size. They are then asked to compute the mean and the range for comparisons with their group's results.

Prompt:

1. Do you think your analysis will be different from your group's data? Why or why not?

Graphs With Cricket Graph

Activity: **Weather Data: Scatterplot**

Students are provided with a database of min., max., and average temperatures for May, 1991. The purpose of this activity is twofold: to provide students with an opportunity to interpret information presented graphically and to encourage students to make predictions about what the graph will look like based on the available data (there is alot of variability in the data from subdegrees to high 20s). This activity allows students to use all the knowledge they acquired from the previous activities.

Prompts:

1. Do you think these scores are similar or do they vary?
2. What might be wrong with the data?
3. Is this the type of data you would expect for the month of May? (i.e., is this typical for this month?)
4. What if you average the values for the years 1990, 1991, & 1992?
5. What does the True Mean show us?

Activity: **World Data: Pie Graph**

Students are provided with a World Bank database which includes variables such as population, education, economy type, literacy, etc. This allowed students to explore a variety of variables and to continue making predictions based on the data. In particular, students were required to examine population by economy type.

Prompt:

1. What population are we interested in here?

Appendix B

Library of Exemplars HyperCard™ Stack: Development

Examples of average and above average performance illustrated in the library of exemplars were selected from videotapes of group projects that were presented by grade eight students in a pilot study conducted in the spring of 1992 (Lajoie et al., 1993). Presentations were time segmented by the experimenter for each of the criteria. Once completed, the experimenter rank ordered average and above average performance for each criterion and then selected the best examples of each criterion. One example of each level of performance was deemed sufficient for the six criteria for the purposes of the current study. Twelve time segments, consisting of video and sound, were then digitized using ScreenPlay™ (Supermac Technology, 1991) and VideoSpigot, a MacIntosh digitizing card. These time segments, or "digitized video clips," were compressed using Adobe Premiere™ (Adobe Systems, 1991). This software also allowed the programmer to edit the movie (i.e. video) and sound as wished. To play the video clips in the HyperCard™ stack, it was necessary to first, specify commands that linked the digitized clips with the appropriate criteria in the HyperCard stack and second, to use QuickTime™ (Apple Computer, 1989) to actually play the video clips. HyperCard was used in this study because of its flexibility in accommodating various types of software. This flexibility opened up the possiblity of using multiple representations for presenting the same information.

Three types of representations were used for the criterion "data analysis" in the library of exemplars stack (see Figure 6 in main text). Text was used to describe the meaning of the criterion. A still image of a video clip was presented on the right hand side of the screen and beneath it were two squares (i.e., "buttons" in HyperCard language) depicting a video camera, each labeled average and above average. Clicking on either allowed students to play the digitized video clips. A close-up screen presenting information more clearly than the video can be accessed by clicking on the close-up button located near the still image (see Figure 7 in main text). This feature was necessary for data analysis and data presentation only. Both close-ups were created by taking a picture of the Mystat™ file containing the data or results using Adobe Photoshop™ (Adobe Systems, 1992). This software allowed the researcher to take a picture of the computer screen as one would take a picture with a camera. This picture was modified using MacPaint® (Claris Corporation, 1987) to exclude irrelevant information and to make the picture smaller so that it could be presented in HyperCard.

Appendix C

Descriptions of Assessment Criteria With Associated Prompts in the Library of Exemplars Stack

1. Quality of Question: 5 points

   State your research question **clearly** so the class will know the purpose of your study.

   What is the difference between these two statements? Do they differ in terms of clarity or specificity? Discuss this with members of your group. You may replay the clips if you are unsure.

2. Data Collection: 10 points

   In order to answer your question you must decide **what data, how much**, and **where** to collect it.

   Notice from the examples that two kinds of data can be gathered. You can collect your *own* data or use information that already *exists* to do your analyses. Aside from this difference, what distinguishes the two groups? How could they both be improved?

3. Data Presentation: 10 points

   Looking at the data gives you an idea of what to expect before you do the analysis. You can illustrate your data either graphically or in raw form. You must **show the information** you have collected, **describe** what it is, and **explain** what it means.

What is missing in the average presentation clip? In what way could this group have presented their data better? Notice that the presenter in the above average clip shows the data graphically and then begins to describe what it is.

4. Data Analysis: 10 points

You can analyze the information that you have gathered by **obtaining statistics** for the mean, median, mode, and range. You must **explain** the results. This demonstrates that you **understand** the significance of the results. You must also consider how your results would change if the study had been done differently.

After looking at the videos, discuss amongst yourselves the differences between the two and why one is better than the other.

5. Presentation Style: 10 points

You will be evaluated according to how well **organized** you are, how well the **graphs and statistics are used** to answer your question, and how **thorough** your explanations are.

Consider how you might want to present your project to the class given the examples you have just seen.

6. Creativity: 5 points

Creativity refers to how **original** you are in presenting your project. As a group, decide upon the *best method* for presenting what you did, how you did it, and what you came up with. Creativity makes your project stand out from those of other groups.

Notice that in the above average group the same data was presented differently by various members of the group whereas the average group presented only one graph. Different types of graphs are just one way of demonstrating creativity. The materials used to present your project may also differ. Can you think of ways to make your project stand out?

Appendix D

Pretest/Posttest

Instructions:

Please answer all of the questions below. Some of the questions will be very difficult but you should write whatever you may think the answer is. It is important that you explain why you got your answer.

1. Suppose you want to find out the favorite fast food restaurant of all the eighth grade students in the province of Québec. How would you find out?

Write a paragraph (point form) describing how you would collect this information. Make sure to include things you would consider such as:

a) what you would ask each person (state the question you would ask)

b) who you would ask and why

c) how many people you would ask and why

d) how you would make a conclusion

Use the back of this page if necessary.

(adapted from Jacobs, 1993, p. 125)

[4] 1.B. Suppose you want to find out whether high school students in the province of Quebec prefer using IBM or Macintosh computers. How would you find out?

Write a paragraph (point form) describing how you would collect this information.

[4] The scenario presented on pretest item #1 was replaced by item #1.B on the posttest in order to avoid practice effects.

Make sure to include things you would consider such as:

a) what you would ask each person (state the question you would ask)

b) who you would ask and why

c) how many people you would ask and why

d) how you would make a conclusion

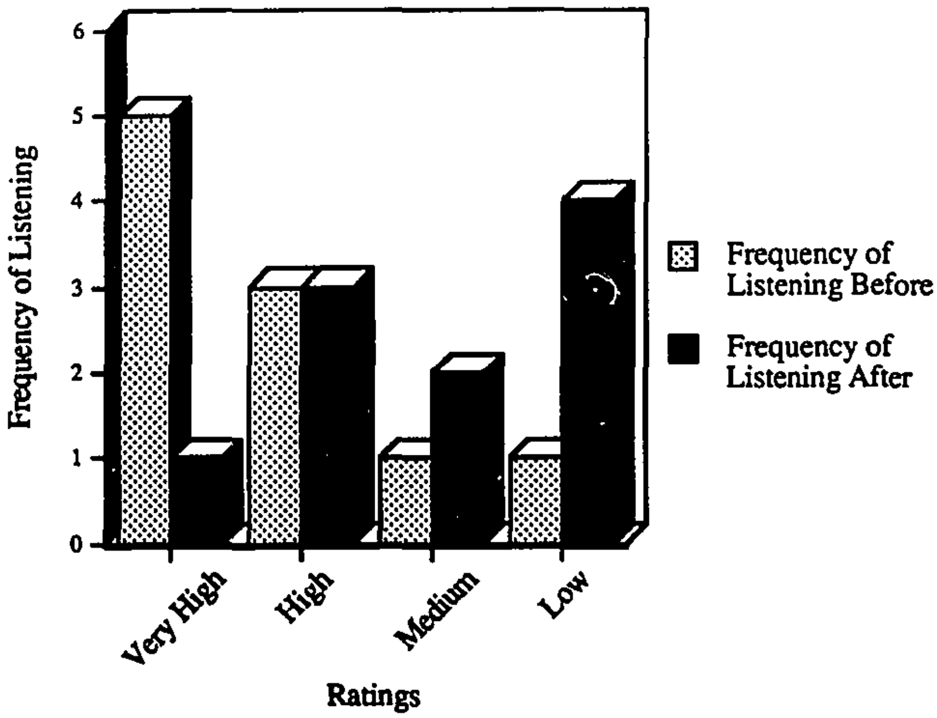
Use the back of this page if necessary.

2. Explain when you would want to use statistics.

3. Explain what is meant by "range" and why is it used in statistics. Why do we need to calculate the range at all?

4. Suppose you are interested in finding out whether Terry Demonté's departure from Chom FM influenced how much people listened to Chom in the morning. The graph shows you the frequency of listening to Chom before Terry left the show and what the ratings are now. Ten people rated the show.

Frequency of Listening to Chom FM Before and After Terry Demonté Left the Station

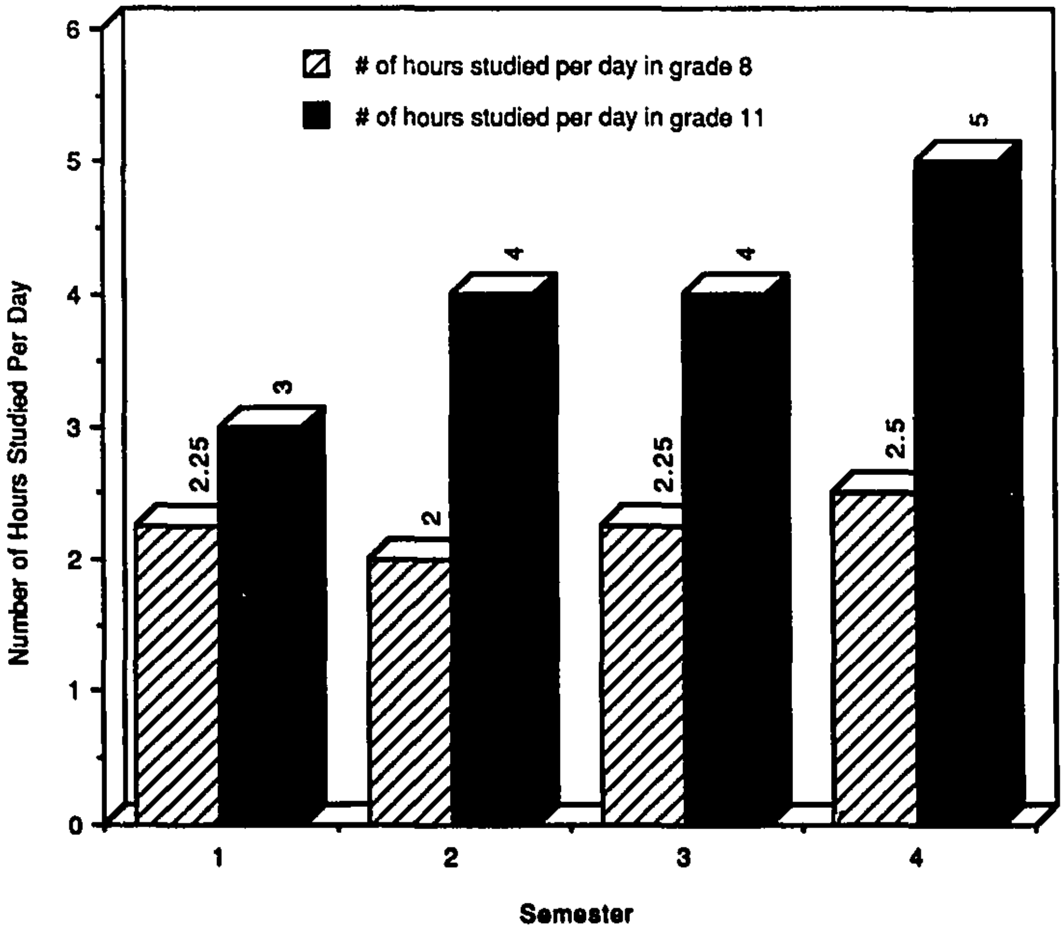


a) What do you conclude when you compare the two ratings? Explain how you arrived at your conclusion.

b) Do you think that this pattern reflects how most people feel about the new morning show? Why or why not?

[5] 4.B. Suppose you were interested in comparing grade 11 and grade 8 students in terms of how many hours they study a day for the 1990-1991 school year. The following graph shows this comparison. Four students in each grade were sampled for the full year.

Number of Hours Grade 8 and Grade 11 Students Study Per Day From First to Fourth Semester in 1991



a) **What do you conclude when you compare the two ratings? Explain how you arrived at your conclusion.**

[5] **Pretest item #4 was replaced by item #4.B on the posttest due to the problematic nature of the original question.**

b) Do you think that this pattern reflects the amount of hours grade 8 and grade 11 students in Québec study per day? Why or why not?

5. Suppose that we have predicted what the temperature will be like for the first 5 days in May. The following table shows the estimated high and low temperatures from May 1st to May 5th, 1993.

	Temperatures	
May 1993	High	Low
1	-1	-2
2	4	0
3	14	9
4	18	14
5	18	15

a) Calculate the "mode" for the high AND low temperatures.

b) Calculate the "median" for both high AND low temperatures.

c) Calculate the "mean" for both high AND low temperatures.

d) Explain the difference between the values obtained in the mean, median, and mode. Which do you think describes the data more accurately and why?

[6]5. B. Suppose that we have predicted what the temperature will be like for the first 5 days in April. The following table shows the estimated high and low temperatures from April 1st to April 5th, 1993.

	Temperatures	
April 1993	High	Low
1	0	-6
2	2	0
3	4	0
4	10	8
6	20	10

a) Calculate the "mode" for the high AND low temperatures.

b) Calculate the "median" for both high AND low temperatures.

c) Calculate the "mean" for both high AND low temperatures.

6. Consider the following scenario.

A researcher, having heard of the recent Ben Johnson scandal, was interested in finding out whether there was a relationship between athletics' use of steroids and type of athletics. Dr. Diaz's research involved Canadian athletes participating in sports such as track and field, football, hockey, and swimming. His prediction that athletes in track and field and in football would use steroids more than athletes in

[6] Numerical values on pretest item #5 were changed to those presented on the posttest to avoid practice effects.

hockey and swimming was confirmed. On this basis, Dr. Diaz concluded that only athletes in track and field and in football use steroids.

a) Identify Dr. Diaz's research question.

b) What is the population in this study?

c) Identify Dr. Diaz's sample.

d) Do you think that Dr. Diaz's conclusion is legitimate? Why or why not?

7. In a small town in Québec, the insurance companies kept records of the claims paid to people because of mistakes that were made by hospitals: damaged equipment, wrong conclusions made by doctors, wrong medecines, etc.

The average for these claims was $69, 000. Half of the claims had a value lower than $8, 000, which means the median was $8, 000.

Give an explanation for the huge gap between the average and the median (adapted from de Lange Jzn. & Verhage, 1992, p. 20)

8. Construct a table which has an outlier in the data. Identify this outlier and explain how it affects the data.

9. Define the following terms:

a) random sample

b) data

10. Suppose that two different agencies took a survey of Brian Mulroney's popularity in the West Island. The scale ranged from most (1) to least popular (10) . Assume 50 people were sampled and that the following results were found:

	Mean Rating	Range
Agency #1	2.5	5.0
Agency #2	8.5	2.0

a) Which agency is more reliable? Why?

b) What factors could have influenced the results?

Appendix E

Order Of Presentation For Pretest/Posttest Items

Pretest Item Number	Corresponding Item Number on Posttest	Statistical Concepts and Procedures
* 1	7	Hypothesis Generation & Sampling (i.e., Sample, Sample Size, Sample Representativeness)
2	4	Statistics
3	3	Measures of Variation (i.e., Range)
* 4	9	Graph Interpretation & Sampling (i.e., Sample Size &Sample Representativeness)
*5	1	Measures of Central Tendency (i.e,., Mode, Median, & Mean)
6	5	Hypothesis Generation, Population, & Sampling (i.e., Sample, Sample Size, & Sample Representativeness)
7	10	Mean & Median
8	6	Outlier
9	2	Randomization & Data
10	8	Range & Sampling (i.e., Sample Size & Sample Representativeness)

Note: (*) refers to items that were replaced in the posttest. All other items are identical except for their location on the posttest.

Appendix F

Scoring Templates for Pretest/Posttest

Question #1 (Pre) **Question #7 (Post)**	**8 points**	**Hypothesis Generation, Sample, Sample Size & Sample Repres.**

Pretest

Supose you want to find out the favorite fast food restaurant of all the eighth grade students in the province of Quebec. How would you find out?

Write a paragraph (point form) describing how you would collect this information. Make sure to include things you would consider such as:

a) hypothesis generation 2 points

1 point = question is general: "What is your favorite restaurant?"

2 points = question is specific: "What is your favorite fastfood restaurant in Québec between..."

b) sample 2 points

1 point = grade 8 students in my class or school

2 points = grade 8 students in Québec (evenly distributed sample).

c) sample size 2 points

1 point = large sample (min.100)

2 points = elaboration (random sampling)

d) conclusion referring to sample representativeness 2 points

1 point = "statement of results" and/or mention of graphs

2 points = mention of generalization of sample to population

Posttest

Supose you want to find out whether high school students in the province of Québec prefer using IBM or Macintosh computers. How would you find out?

Write a paragraph (point form) describing how you would collect this information. Make sure to include things you would consider such as:

a) hypothesis generation 2 points

1 point = question is general "What type of computer do you prefer?"

2 points = question is specific "What is your favorite type of computer: IBM or Macintosh?"

b) sample 2 points

1 point = high school students in my class or school

2 points = high school students in Québec (evenly distributed sample).

c) sample size 2 points

1 point = large sample (min.100)

2 points = elaboration (random sampling)

d) conclusion referring to sample representativeness 2 points

1 point = "statement of results" and/or mention of graphs

2 points = mention of generalization of sample to population

Question #2 (Pre) **2 points** **Statistics**
Question #4 (Post)

Explain when you would want to use statistics.

1 point = if say that the purpose is to describe OR to predict or mention calculations

2 points = if they say the purpose it to describe AND predict

Question #3 (Pre) **Question #3 (Post)**	**2 points**	**Range**

Explain what is meant by "range" and why it is used in statistics. Why do we need to calculate the range at all?

1 point = definition of range

1 point = to indicate variability -- supplements mean or median or mode

Question #4 (Pre) **Question #9 (Post)**	**5 points**	**Graph Interpretation, Sample Size, & Sample Representativeness**

Pretest:

Graph representing the frequency of listening to Chom FM before and after Terry Demonté left the station.

a) Interpreting the graph: Comparing two ratings. Explanation of how conclusion was made. 2 points

1 point = ratings before are higher than ratings after therefore ratings drop (qualitative)

1 point = drop is largely accounted by the shift from those in high to the low rating (elaboration)

b) Sample Size & Sample Representativeness: Statement of whether this pattern reflects how most people feel about the new morning show and explanation of why this is or is not the case. 3 points

1 point = No as answer

1 point = explanation- sample size issue (10 is an insufficient #)

1 point = explanation- sample representativeness issue (depends on who you ask)

Posttest

Graph of Number of Hours Grade 8 and Grade 11 Students Study Per Day from First to Fourth Semester in the 1990-1991 School Year .

a) Interpreting the graph: Comparing two ratings. Explanation of how conclusion was made. 2 points

1 point = grade 11 students study more than grade 8 students

1 point = grade 11 students steadily study more as the semester progresses whereas grade 8 students study habits do not change from semester to semester (i.e. 2 hours).

b) Sample Size & Sample Representativeness: Statementof whether this pattern reflects the amount of hours grade 8 and grade 11 students in Québec study per day and an explanation of why the conclusion was made. 3 points

1 point = No as an answer

1 point = explanation- sample size (4 is an insufficient #)

1 point = representativeness issues (depends on who you ask and which school - i.e. private versus public school).

Question #5 (Pre) **Question #1 (Post)**	**10 points**	**Mode, Median, &** **Mean**

Weather data (high and low temperatures for May 1991) is presented in a table.

a) Mode for high and low temperatures 2 points

1 point for each calculation

Answer: High = 18 Low = none

b) Median for high and low temperatures 2 points

1 point for each calculation

Answer: High = 14 Low = 9

c) Mean for high and low temperatures 2 points

1 point for each calculation

Answer: High = 9.83 Low = 7

d) Explanation of differences between measures of central tendency and selection of better measure. 4 points

Students can differentiate between measures in terms of definitions or in terms of the data presented in the problem.

1 point = *means* are affected by extreme scores (May 1st)

1 point = some data do not have a *mode* (no value for low temp on May 1st)

2 points = *median*

1 point = say median is better without an explanation

1 point = say median with appropriate explanation (refers to how median is least affected by extreme or low scores)

Posttest

Weather data is presented in a table. Calculate measures of central tendency. Comparison of these measures.

a) Mode for high and low temperatures 2 points

1 point for each calculation

Answer: High = none Low = 0

b) Median for high and low temperatures 2 points

1 point for each calculation

Answer: High = 4 Low = 0

c) Mean for high and low temperatures 2 points

1 point for each calculation

Answer: High = 7.2 Low = 2.4

d) Explanation of differences between measures of central tendency and selection of better measure. 4 points

Students can differentiate between measures in terms of definitions or in terms of the data presented in the problem.

1 point = some data do not have a *mode* (no value for high temperature on April 1st)

1 point = *median* not truly representative given temperature for April 5 (would be if excluded this value)

2 points = *mean*

1 point = say mean is better without an explanation

1 point = say mean with appropriate explanation (despite being influenced by extreme score-April 5 high & April 1 low- it is still most appropriate. If excluded this value, median would be best)

Question #6 (Pre) **Question #5 (Post)**	**10 points**	**Hypothesis Identificiation, Popn, Sample, Sample Size, & Sample Rep.**

Ben Johnson example -- samples and populations

a) hypothesis identification 2 points

Answer: whether there is a relationship between athletics' use of steroids and type of athletics.

b) population 3 points

1 point: athletes

2 points: athletes in the world

3 points: athletes in the world participating in track & field etc.

c) sample 2 points

1 point = Canadian athletes

1 point = Participating in track and field, football, hockey, and swimming

d) sample size & sample representativenss: evaluation of conclusion presented in scenario 3 points

1 point = not legitimate

1 point = sample limited to a few sports (did not look at full range)

1 point = sample limited to Canadian athletes, not generalizable to all athletes, particularly with a sample of 50.

Question #7 (Pre) **2 points** **Mean & Median**
Question #10 (Post)

Insurance company example -- difference between mean and median within this context.

1 point = definition of the mean and/or median

e.g. The mean uses all scores but the median is the number that falls in the middle.

1 point = explanation of the huge gap in the mean and median values

e.g. Thus, the highest claims are above the median. Such extreme scores do not affect the median but do affect the mean.

Question #8 (Pre) **Question #6 (Post)**	**2 points**	**Outlier**

Construct a table showing an outlier. Indicate which is the outlier and explain how it influences the data.

1 point = table with outlier

1 point = explain influence on the mean in particular

Question #9 (Pre) **Question #2 (Post)**	**4 points**	**Randomization & Data**

Definition of concepts

a) random sample 2 points

1 point = describe what sample means

e.g. pick a portion at random with no example

1 point = specify randomness through examples or definition

i.e. link population to sample - sample looks like the population

b) data 2 points

2 points = data is information

Question #10 (Pre) **Question #8 (Post)**	**4 points**	**Range, Sample Size & Sample Rep.**

Brian Mulroney example -- role of range and factors affecting conclusions

Pretest

a) range: comparison of agencies 2 points

1 point = agency #2 is more reliable

1 point = reason: the mean is higher and the range smaller for agency #2 (i.e., little variability)

b) sample size & sample representativeness: identification of extraneous variables-conclusions 3 points

1 point = interviewing techniques or type of questions asked may differ from agency to agency (i.e. may not have been asking exactly same thing)

2 points = sampling issues

1 point = sample size

1 point = sample representativeness

Posttest

a) range: comparison of agencies 2 points

1 point = agency #2 is more reliable

1 point = reason: the means are the same for both agencies but the range smaller for agency #2 (i.e., little variability)

b) sample size & sample representativeness: identification of extraneous variables-conclusions 2 points

1 point = interviewing techniques or type of questions asked may differ from agency to agency (i.e. may not have been asking exactly same thing)

2 points = sampling issues

1 point = sample size

1 point = sample representativeness

Total Points: 50 points

Appendix G

Homework Assignment #1

Instructions:

Please read the following paragraph carefully. You will be required to answer a set of questions about this paragraph once you are finished. This may be difficult for you right now, but write what you think would be the best answer to these questions and explain why.

Dr. Aloe, a researcher, is interested in finding out whether students learn mathematics better with computers than without computers. Twenty grade 6 students attending a public high school were randomly assigned to two groups: 10 students learned mathematics using computers, whereas the other 10 students did not use computers. Both groups learned the same material and were given the same test. Comparison between the two groups are made on the basis of test scores (maximum score is 50).

The data for the two groups is the following:

Computers	No Computers
40	35
35	30
42	37
43	40
47	39
41	39
45	42
44	38
40	40
43	36

1. What is Dr. Aloe's hypothesis?

2. Explain how Dr. Aloe gathered his data.

3. Are the two groups (i.e. computer vs. no computer) different? Why or why not?

4. Can you compute some differences? If so, show how you would do this.

5. Could you show results (i.e. two groups are different or the same) in a graph? Explain how you would do this or draw a graph showing this.

Appendix H

Homework Assignment #2

Suppose that after this one-week tutorial you become a leading expert in statistics at Lindsay Place High School. Students going through the same tutorial next year will therefore come to you for advice. What **examples** would you give to help these students **come up with** ideas for doing their own study? (Please try to use examples other than those discussed in class).

Once you have written down some examples, please explain in a paragraph or two the 4 steps involved in doing an experiment and how these relate to concepts you have learned (e.g. data, population, sample, data , mean, median, mode, and range).

Appendix I

Scoring Template For Homework Assignment #1

Question #1 **1 point** **Research Question**

What is Dr. Aloe's hypothesis? *Population, Sample*

1 point = To determine whether students learn mathematics better with computers than without computers

Question #2 **2 points** **Data Collection**

Explain how Dr. Aloe gathered his data *Population, Sample, Randomization*

1 point = If students simply reproduce the problem statement found in the text

i.e., 20 grade 6 students attending a public high school were randomly assigned to 2 groups.

2 points = If students mention statistical terms, going beyond the problem statement

i.e., collected a sample of 20 grade 6 students from a population of students.

Question #3 **3 points** **Data Analysis & Sampling Considerations**

Are the 2 groups (i.e. computer vs. no computer) different? Why or why not? *Data, Mean, Sample size & representativeness.*

1 point = If students write that the 2 groups are different

2 points = If students say the 2 groups are different because of one group worked with computers and the other didn't

3 points = If students say that the 2 groups are different based on descriptive statistics

i.e., "the mean was higher for the computer group therefore they learned best."

Question #4 **2 points** **Data Analysis**

Can you compute some differences? If so, show how you would do this. *Mean, Range*

1 point = If students compute the average or say how they would do so

2 points = If students compute the average and the range.

	Mean	Range
With Computers	42.0	12
Without Computers	37.6	12

Question # 5 **2 points** **Data Presentation**

Could you show results (i.e., 2 groups are different or the same) in a graph? Explain how you would do this or draw a graph showing this. *Data,Statistics, Graphs*

1 point = If students explain how to draw a graph (or) draw a graph showing data for each score rather than average (or) if the graph is not clear about the differences (e.g., axes are labelled inappropriately)

2 points = If students draw a graph that illustrates the difference between the 2 groups in terms of average and not individual scores (i.e., 2 bars in the column graph rather than 20)

Total: **10 Points**

Appendix J

Scoring Template For Homework Assignment #2

Ideas/Examples **1 point** **Creativity**

0.5 point = If the student gives an example presented in class during instruction or group projects

1 point = If the student gives a novel example

The 4 Steps Involved in Conducting an Experiment **9 points** **Research Question**
Data Collection
Data Analysis
Data Presentation

- *Research Question* *2 points* *Population, Sample*

1 point = If the student says anything pertaining to hypothesis generation

2 points = If the student mentions that hypotheses are made about populations and that samples are used to make inferences about this population

- *Data Collection* *3 points* *Data, Sample, Sample size, Randomization & Representativenss*

1 point = If the student mentions the fact that data needs to be collected in order to try to answer the research question

2 points = If the student mentions that the data is gathered from a sample which represents the population

3 points = If the student mentions issues of sampling such as sample size, randomization, and sample representativeness

- *Data Analysis* *3 points* *Mean, Mode, Median, & Range*

1 point = If the student says that data must be analyzed

2 points = If the student says that means, modes, medians, & ranges can be used to analyze data

3 points = If the student provides a description or explains these terms and/or comments are made about the influence of outliers or extreme scores are made

- *Data Presentation* *1 point* *Data, Statistics, Graphs*

1 point = If the student mentions the use of graphs or charts for displaying the results.

Total: **10 points**

Appendix K

Knowledge, Reasoning, Planning, and Reflection Prompts in Structured Group Journals

Knowledge Prompts

Definition of Concepts

Statistical concepts were taught in a 10-15 minute lecture and during a statistics tutorial which required that students work in groups. Group journals were structured to encourage students to document the meaning of these concepts during statistical activities. Of interest was the way in which groups formulated their declarative knowledge.

Prompts:

1. What is data?
2. What is statistics?
3. What does population refer to?
4. What is a sample?
5. What does randomization refer to?
6. Name three statistics that can be used when analyzing data. What kind of statistics are these?
7. What is a mean? How is it calculated?
8. What is an outlier?
9. What does mode refer to?
10. What is meant by the term median?
11. What is "range?" How is the range calculated?

Explanation of Concepts Groups were given the opportunity to elaborate on their understanding of various statistical concepts through prompts that encouraged explanation. This allowed groups to develop a conceptual understanding of statistics.

Prompts:

1. What five steps are involved in conducting a statistics experiment?
2. Give 5 examples of how statistics is used in the media (e.g., newspaper).
3. What are the two uses of statistics?
4. What are the two ways in which data and statistics can be presented?
5. Give an example of a statistic.
6. Why are means important?
7. What types of scores are problematic for means (2)?
8. What are the pitfalls of the median and mode?
9. Why do we need to calculate the range?
10. Why are graphs used in statistics?
11. Does your group think that looking at the data is important? Why or why not?

Reasoning Prompts

Reasoning Based on Data and Graphs in Tutorial The prompts found in the structured group journals were identical to those embedded within the statistics tutorial. Such prompts encouraged groups to think and reason about the data and graphs they were working with on the tutorial.

Prompts: See Appendix A.

Reasoning Based on Project Data

Groups of students were required to conduct their own study and were encouraged to write their ideas and their plans regarding their projects and presentations in their journals.

Prompt:

1. Of the graphs that you have tried, which ones does your group think will be better?

Planning Prompts

Generation of Ideas

The purpose of incorporating planning prompts in group journals was to encourage students to generate ideas and to use these as frameworks for learning and applying their knowledge of statistical concepts and procedures. Groups were encouraged to begin planning their projects from the outset and were reminded of their task throughout the structured journal. The journal emphasized planning particularly towards the end of the tutorial when groups were learning how to construct graphs.

Prompts:

1. Do your group have any ideas for a possible project? If so, list them here.
2. Does your group think they will use the types of graphs you learned about in the tutorial for your projects?
3. List some ideas your members might have for your group's project. What was your first idea?

4. What is your group's research question?
5. Who are you going to ask to participate in your study?
6. How many people does your group need to make sure your sample is representative?
7. How is your group going to demonstrate the data you have collected?
8. What statistics does your group want to do?
9. What analyses has your group tried?
10. What graphs has your group tried?
11. How are you going to present your project to the class?
12. What statistics has your group decided to do? Why?
13. What graphs has your group decided to do? Why?

Reflection Prompts

Evaluation of Understanding

Groups were encouraged to ask questions and to evaluate their own understanding through prompts found within the group journal. At the end of each section, groups were given the opportunity to document their thoughts and ideas that were generated as a result of their experience.

Prompts:

1. What questions does your group have?
2. What does your group not understand?
3. What does your group understand?
4. List any thoughts that your group might have about the day's activities.
5. What two types of graphs did your group learn in the tutorial? (say a little bit about them).

6. What future research could your group do based on the results of your study?
7. What suggests would your group have for students participating in this study next year?
8. If you could create a problem for next year's students, what would it be?

Appendix L

Mini-Course Evaluation

1. Did you like this mini-course ?

(adapted from Jacobs, 1993, p. 157)

1	2	3	4	5
I liked it alot		It was OK		I did not like it at all

2. What did you like about this course?

(adapted from Jacobs, 1993, p. 157)

3. What did you dislike about this mini-course?

(adapted from Jacobs, 1993, p. 157)

4. What things during the McGill project helped you understand statistics the most?

5. Did the your classmate's presentations help you understand statistics a little bit better? If so, name 3 things that you learned from just watching your friends present their statistics project.

6. Did you keep your first idea for your group project? If you did not, why did you change it?

7. Would you change your group projects now that everyone has presented? If so, why would you change it and how would you do it?

8. What suggestions do you have to make this mini-course better?

(adapted from Jacobs, 1993, p. 157)