Statistics Education Research Journal

Volume 6 Number 2 November 2007

International Association for Statistical Education
http://www.stat.auckland.ac.nz/~iase

International Statistical Institute
http://isi.cbs.nl/

STATISTICS EDUCATION RESEARCH JOURNAL

The *Statistics Education Research Journal* (*SERJ*) is a peer-reviewed electronic journal of the International Association for Statistical Education (IASE) and the International Statistical Institute (ISI). *SERJ* is published twice a year and is free.

SERJ aims to advance research-based knowledge that can help to improve the teaching, learning, and understanding of statistics or probability at all educational levels and in both formal (classroom-based) and informal (out-of-classroom) contexts. Such research may examine, for example, cognitive, motivational, attitudinal, curricular, teaching-related, technology-related, organizational, or societal factors and processes that are related to the development and understanding of stochastic knowledge. In addition, research may focus on how people use or apply statistical and probabilistic information and ideas, broadly viewed.

The *Journal* encourages the submission of quality papers related to the above goals, such as reports of original research (both quantitative and qualitative), integrative and critical reviews of research literature, analyses of research-based theoretical and methodological models, and other types of papers described in full in the Guidelines for Authors. All papers are reviewed internally by an Associate Editor or Editor, and are blind-reviewed by at least two external referees. Contributions in English are recommended. Contributions in French and Spanish will also be considered. A submitted paper must not have been published before or be under consideration for publication elsewhere.

Further information and guidelines for authors are available at: http://www.stat.auckland.ac.nz/serj

Submissions

Manuscripts must be submitted by email, as an attached Word document, to co-editor Tom Short <tshort@iup.edu>. Submitted manuscripts should be produced using the Template file and in accordance with details in the Guidelines for Authors on the Journal's Web page: http://www.stat.auckland.ac.nz/serj

Publication: IASE/ISI, Voorburg, The Netherlands
Technical Production: California Polytechnic State University, San Luis Obispo, California, United States of America
Web hosting and technical support: Department of Statistics, University of Auckland, New Zealand

ISSN: 1570-1824

SERJ EDITORIAL BOARD

TABLE OF CONTENTS

Editorial 2

Dustin L. Jones and James E. Tarr
An Examination of the Levels of Cognitive Demand Required by Probability Tasks in Middle Grade Mathematics Textbooks 4

Robert delMas, Joan Garfield, Ann Ooms, and Beth Chance
Assessing Students' Conceptual Understanding After a First Course in Statistics 28

Scott R. Evans, Rui Wang, Tzu-Min Yeh, Jeff Anderson, Rammy Haija, Paul Madoc McBratney-Owen, Lynne Peeples, Subir Sinha, Vanessa Xanthakis, Natasa Rajcic, and Jiameng Zhang
Evaluation of Distance Learning in an "Introduction to Biostatistics" Class: A Case Study 59

Dirk T. Tempelaar, Sybrand Schim van der Loeff, and Wim Gijselaers
A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance 78

Past IASE Conferences 103

Forthcoming IASE Conferences 105

Other Past Conferences 110

Other Forthcoming Conferences 111

Statistics Education Research Journal Referees 112

EDITORIAL

This last August I was fortunate enough to be able to attend several meetings, including SRTL-5 (the fifth forum on Statistical Reasoning, Thinking, and Literacy, at the University of Warwick, UK), the IASE Satellite meeting on Assessing Student Learning in Statistics (Guimarães, Portugal), and ISI-56, the biannual meeting of the International Statistical Institute (Lisbon, Portugal). Information about all of them appears in the "Past IASE Conferences" section at the end of this issue.

It was exciting to chat informally and hear presentations regarding a very wide range of studies, projects, and professional activities related to statistics education. Clearly, the international community interested in research on the learning, teaching, and understanding of statistics and probability, is growing and diversifying. From the many topics I came across, I would like to briefly highlight one that deserves special mentioning in the context of a research journal such as *SERJ*, related to the types of research data and types of evidence we encounter, and their implications for research publishing and for teaching/learning.

We often speak of "quantitative research" versus "qualitative research." Although it is recognized that both types are needed in research of an educational nature, sometimes we see researchers leaning towards one or the other. There is a somewhat tenuous relationship between quantitative and qualitative research in an area whose subject matter, statistics, is based on quantitative information, and where some of the researchers and teachers (as well as manuscript referees…) are mainly trained in quantitative methods.

However, I have now come across a number of situations where neither of these two traditional labels is sufficient, and perhaps we should refer to a third (hybrid?) kind, "Dynamic data." The need to rethink the traditional division of research into quantitative and qualitative became obvious to me this summer when listening to reports about classroom activities and studies where learners and teachers used dynamic software such as Fathom, Tinkerplots, or interactive applets such as probability simulators. In such and related cases, the data being collected by researchers (i.e., information about what students did, what they looked at, and how they thought during an activity or interpreted the results) was more complex than ever before, and sometimes quite slippery. The data accumulated over time and involved a dynamically changing mix of elements such as utterances and conversations among students or among students and teacher, different types of graphical displays, multiple "what if" trials with different aggregations or data views that the students looked at in the course of their work, results of trying different kinds of simulations, and more.

Of course, the need to collect, describe and integrate data from multiple sources, both quantitative and qualitative, has existed before the emergence of dynamic software. However, listening to reports from different studies, it became apparent that researchers are challenged by the need to capture and describe the additional fast-changing and multi-faceted data generated when dynamic software is an inherent part of the teaching/learning environment and when students are given enough time to use it in an exploratory manner. The nature of what students look at, work with, refer to, or think about is becoming more complex and harder to document, as it rapidly changes over time. Of course, all these realities place additional burden and present new demands to teachers working in a "dynamic data" environment, and have implications for the forms of needed assessments. Further, researchers need new tools, methods, terminology, or conceptualizations in order to analyze and interpret such data, and probably so do teachers. The need to report in a concise and coherent manner what transpires in a teaching/learning episode involving dynamic data in turn presents new challenges to researchers trying to write a compact manuscript for publication in a journal such as *SERJ*.

It follows that new technology-based developments offer brave new worlds for educators, learners, and researchers alike, and promise to make learning more fun, interesting, and deeper in nature. Yet, such developments also make life more complex for all involved. Certainly, as more researchers would want to report the results of research using “dynamic data” as described above, research journals such as *SERJ* may need to consider “dynamic reporting” of data, such as in the form of links within documents to mini-videos or dynamic screen-shots so that readers can appreciate the nature of the information being analyzed and reported by researchers.

While the observations and ideas presented above are tentative in nature, certainly they may cause us to think where our field is moving. Next year, in 2008, several important meetings will take place where such and related developments can be further examined and discussed, and they are listed in the “Forthcoming Conferences” section in this issue. I refer in particular to two Topic Study Groups, #13 and #14, to be held as part of ICME-11 (International Congress on Mathematical Education), which will deal with research and development in the teaching and learning of probability, and of statistics, respectively. In addition, prior to ICME, the special “Joint ICMI/IASE study on statistics education in school mathematics” will be another forum where tensions and responsibilities emerging due to new technologies can be further explored.

This issue of *SERJ* is the last that I will be co-editing, having reached the end of my four-year term. It is a pleasure for Tom and me to announce that Peter Petocz was appointed as co-editor for *SERJ* for the years 2008-2011 by the IASE Executive Committee, following the unanimous recommendation of the IASE search committee. Peter is Associate Professor in the Department of Statistics at Macquarie University, Australia. He is a very innovative and effective statistics educator, and also an accomplished researcher who has published on pedagogical issues in statistics and mathematics education. Peter will soon begin working with Tom Short, who continues as co-editor through 2009, and I wish both of them a good time ahead.

In closing, I would like to express my gratitude to the many dedicated members of the *SERJ* Editorial Board and to the journal’s many referees who continue to invest time and effort in helping to improve research publishing and contribute advice and support to authors and educators alike. The growth *SERJ* has experienced over the last four years has been also helped by the support and understanding of the IASE Executive committee and its former and current presidents. All this goes to show that a journal such as *SERJ* develops in a dynamic environment that is sometimes slippery, yet full of promise. I am certain that the new editorial team will continue to find ways to maintain quality in published manuscripts, yet at the same time enable *SERJ* readers to benefit from new opportunities for developing research-based knowledge in our evolving field.

IDDO GAL, for TOM SHORT

AN EXAMINATION OF THE LEVELS OF COGITIVE DEMAND REQUIRED BY PROBABILITY TASKS IN MIDDLE GRADES MATHEMATICS TEXTBOOKS

DUSTIN L. JONES
Sam Houston State University
dljones@shsu.edu

JAMES E. TARR
University of Missouri – Columbia
tarrj@missouri.edu

ABSTRACT

We analyze probability content within middle grades (6, 7, and 8) mathematics textbooks from a historical perspective. Two series, one popular and the other alternative, from four recent eras of mathematics education (New Math, Back to Basics, Problem Solving, and Standards) were analyzed using the Mathematical Tasks Framework (Stein, Smith, Henningsen, & Silver, 2000). Standards-era textbook series devoted significantly more attention to probability than other series; more than half of all tasks analyzed were located in Standards-era textbooks. More than 85% of tasks for six series required low levels of cognitive demand, whereas the majority of tasks in the alternative series from the Standards era required high levels of cognitive demand. Recommendations for future research are offered.

Keywords: *Probability; Curriculum; Mathematics textbook content analysis; Mathematical tasks; Cognitive demands; Middle grades mathematics*

1. INTRODUCTION

1.1. THE EMERGENCE OF PROBABILTY IN SCHOOL MATHEMATICS

Consumers and citizens in today's information-rich society need to have an understanding of probability. Shaughnessy (1992) stated, "There is perhaps no other branch of the mathematical sciences that is as important for *all* students, college bound or not, as probability and statistics" (p. 466, emphasis in original). Despite the importance of probability and statistics, many children and adults hold misconceptions about probability (Garfield & Ahlgren, 1988; Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993). In fact, Garfield and Ahlgren (1988) stated that "inappropriate reasoning [in probability and statistics] is…widespread and persistent…and similar at all age levels" (p. 52).

Instruction in probability should provide experiences in which students are allowed to confront their misconceptions and develop understandings based on mathematical reasoning (Garfield & Ahlgren, 1988; Konold, 1983, 1989; Shaughnessy, 2003). Due to the widespread nature of probabilistic misconceptions among adults, such instruction may not have occurred for all students in recent decades. Perhaps probability topics were not present in textbooks, or perhaps these topics were present in textbooks but omitted from instruction (Carpenter, Corbitt, Kepner, Lindquist, & Reys, 1981; Shaughnessy, 1992).

Statistics Education Research Journal, 6(2), 4-27, http://www.stat.auckland.ac.nz/serj

Teachers may omit topics for a number of reasons, including a lack of preparation to teach topics from probability due to their own lack of experience or misconceptions (Conference Board of the Mathematical Sciences, 2001), or a subsequent lack of confidence in their ability to teach such topics. In some cases, a teacher's interpretation of and orientation to the curriculum may constrain how what is printed in the textbook is communicated to the class (Remillard & Bryans, 2004). Moreover, teachers' interpretations of the textbook may be in opposition to the intentions of the authors (Lloyd, 1999; Lloyd & Behm, 2005). In such cases, a teacher may use all of the probability tasks in an investigation-oriented textbook, but present these tasks in a traditional manner by providing students with explicit rules, formulas, and repetitive practice problems. Alternatively, teachers may not have time to teach all of the material present in the textbook, and omit probability lessons simply for a lack of sufficient time. If probability topics are taught, the lessons as presented in textbooks may not sufficiently address students' misconceptions.

Over the past several decades, probability has emerged as an important topic for all students to learn, particularly those in the middle grades. Due to the growing emphasis on the topic of probability in recommendations from professional organizations, one might reasonably expect to observe changes in textbooks. Unfortunately, there have not been any systematic examinations of the composition of textbooks as they have evolved over time, particularly in relation to probability.

1.2. TEXTBOOK USE AND LEVEL OF COGNITIVE DEMAND

Textbooks are common elements in classrooms throughout the world, and are ubiquitous in mathematics classrooms in the United States. Textbooks are present not only in classrooms, they are also frequently used by teachers and students, and influence the instructional decisions that teachers make on a daily basis (Robitaille & Travers, 1992; Tyson-Bernstein & Woodward, 1991). Recent studies have revealed that most middle-grades (grades 6-8) mathematics teachers use most of the textbook most of the time (Grouws & Smith, 2000; Weiss, Banilower, McMahon, & Smith, 2001). Grouws and Smith (2000) observed that the mathematics teachers of three fourths of the eighth grade students involved in the 1996 National Assessment of Educational Progress (NAEP) reported using their textbook on a daily basis. Weiss et al. (2001) found that two thirds of middle-grades mathematics teachers "cover" at least three fourths of the textbook each year. These findings tend to agree with results of research on students' use of mathematics textbooks as well. In the 2000 administration of the NAEP, 72% of participating eighth graders stated that they did mathematics problems from a textbook every day (Braswell et al., 2001).

After analyzing the levels of cognitive demand of mathematical tasks, QUASAR [Quantitative Understanding: Amplifying Student Achievement and Reasoning] project researchers noted that students "need opportunities on a regular basis to engage with tasks that lead to deeper, more generative understandings about the nature of mathematical concepts, processes, and relationships" (Stein, Smith, Henningsen, & Silver, 2000, p. 15). They also found that teachers implementing tasks with high levels of cognitive demand rarely selected tasks from commercial textbook series (Stein, Grover, & Henningsen, 1996).

2. RESEARCH OBJECTIVES

Because textbooks have a marked influence on what is taught in mathematics classrooms, it is important to investigate curricular materials that many mathematics teachers use and the potential of such resources to impact students' opportunities to learn probability. Accordingly, the research reported in this article addressed the following research questions: What is the nature of the treatment of probability topics in middle grades mathematics textbooks? How has the nature of the treatment of probability changed over the past 50 years and across popular textbooks series and alternative (or innovative) textbook series? More specifically, what levels of cognitive demand are required by tasks and activities related to probability, and what are the trends in the required level of cognitive demand over the past 50 years?

Heretofore there has not been any systematic review of the content of textbooks over time. Thus, our study is intended to highlight any differences that have come about through different eras of mathematics education, and how these differences coincide with the contemporary recommendations for the inclusion of probability in the school mathematics curriculum. Moreover, it is our goal to reveal the degree to which textbooks have maintained the status quo in terms of the content and level of cognitive demand required by tasks.

3. THEORETICAL CONSIDERATIONS

3.1. RECENT TEXTBOOK CONTENT ANALYSES

In a pivotal study, Project 2061 (American Association for the Advancement of Science [AAAS], 2000) analyzed thirteen contemporary mathematics textbook series written specifically for middle grades students. Their sample of textbooks included four series developed with support from the National Science Foundation; the other nine series were popular textbooks from the late 1990s. They evaluated each series according to a set of benchmarks related to the core content that should be present in middle grades mathematics instruction: number concepts, number skills, geometry concepts, geometry skills, algebra graph concepts, and algebra graph skills. It is important to note that topics from probability were excluded from their analysis. The research team examined the student and teacher editions of each textbook, specifically attending to lessons that dealt with their selected benchmarks. Each series was rated as having most, partial, or minimal content according to each benchmark. The research team found that only four of the series addressed four or more benchmarks in depth, and no series sufficiently addressed all of the benchmarks. Finally, in terms of quality, none of the popular textbooks were among the best rated.

Valverde, Bianchi, Wolfe, Schmidt, and Houang (2002) also analyzed the content of the textbooks in their sample according to the characteristics of lessons. These characteristics included the primary nature of lessons (concrete and pictorial vs. textual and symbolic), components of the lesson, and student performance expectations. To measure textbook lessons along these dimensions, the researchers divided lessons into blocks, "classified according to whether they constituted narrative or graphical elements; exercise or question sets; worked examples; or activities" (p. 141). The research team analyzed these blocks according to the mathematical topics that were addressed. Results related to the treatment of probability were not reported because topics from this branch of mathematics were not present in many of the textbooks. The researchers also identified the student performance expectations for each block. This analysis revealed that mathematics "textbooks across all populations were mostly made up of exercises and question sets" (p. 143). Additionally, over the three grade levels, the amount of narrative and worked examples increased, whereas the number of activities decreased. Furthermore, "the most common expectation for student performance was that they read and understand, recognize or recall or that they use individual mathematical notations, facts or objects. This is followed . . . by the use of routine mathematical procedures" (p. 128). In order to describe the characteristics of probability tasks in middle grades mathematics textbooks, we utilized a methodology similar to that described by Valverde et al. We identified all of the probability tasks within a textbook, and coded each task with the level of cognitive demand required by the task.

In an effort to provide a tool for comparing the intended, enacted, and assessed curricula, Porter (2006) developed two-dimensional languages to describe the content of the mathematics curriculum. This two-dimensional language can be presented in a rectangular matrix with *topics* as rows and *cognitive demands* (sometimes called *performance goals* or *performance expectations*) as columns. Topics are content distinctions such as "add whole numbers" or "point slope form of a line." Cognitive demands distinguish memorizing; performing procedures; communicating understanding of concepts; solving non-routine problems; and conjecturing, generalizing, and proving. Our research utilizes methodology similar to that which Porter has described, in that we examined the content of textbooks in terms of topics and levels of cognitive demand.

Very recently, the National Research Council [NRC] (2004) issued a key report evaluating the evidence regarding the effectiveness of K-12 mathematics textbooks. The authors devoted an entire chapter to content analysis, and provided descriptions of the methodology and results on several recently published textbook evaluations in the form of content analyses (e.g., AAAS, 2000), as well as unpublished reports available on the world wide web (e.g., Adams et al., 2000; Clopton, McKeown, McKeown, & Clopton, 1999a, 1999b, 1999c; Robinson & Robinson, 1996). Consistent with the recommendations of the NRC, we address the depth of mathematical inquiry and reasoning of probability tasks in textbooks by rating these tasks according to their level of cognitive demand. Our analysis of the levels of cognitive demand required by tasks further provides insight into the engagement, timeliness, and support for diversity provided in each textbook. Textbooks containing tasks that predominately require lower levels of cognitive demand may not support student learning because students are rarely asked to grapple with difficult situations.

3.2. DEVELOPMENT OF THE MATHEMATICAL TASKS FRAMEWORK

Research on tasks as the primary unit of instruction and learning began in the late 1970s and early 1980s. During that time, Doyle (1983) provided the groundwork that would become influential in the work of the QUASAR research team. Doyle described students' work in terms of academic tasks. He used this term to focus on the following:

> (a) the products students are to formulate, such as an original essay or answers to a set of test questions; (b) the operations that are to be used to generate the product, such as memorizing a list of words or classifying examples of a concept; and (c) the "givens" or resources available to students while they are generating a product, such as a model of a finished essay supplied by the teacher or a fellow student. Academic tasks, in other words, are defined by the answers students are required to produce and the routes that can be used to obtain these answers. (p. 161)

In later work, Doyle (1988) added a fourth component of academic tasks as “the importance of the task in the overall work system of the class” (p. 169). It should be noted here that Doyle considered individual questions, exercises, or problems as distinct academic tasks. He defined four general categories of academic tasks: memory tasks, procedural or routine tasks, comprehension or understanding tasks, and opinion tasks (Doyle, 1983). He argued that each of these categories varied in terms of the cognitive operations required to successfully complete tasks contained therein.

The research on academic tasks mentioned above provided a theoretical foundation for the Mathematical Tasks Framework developed by the QUASAR Project team (Smith & Stein, 1998; Stein et al., 1996; Stein & Smith, 1998; Stein et al., 2000). The Mathematical Tasks Framework represents the relationship between student learning and three phases of task implementation. In this model, tasks are first represented in curricular materials, then set up by teachers, and finally implemented by students in the classroom. This framework is particularly useful for our study, because it gives specific attention to tasks as they are present in textbooks.

This model further delineates four levels of cognitive demand for tasks: lower-level demands of Memorization and Procedures without Connections, and higher-level demands of Procedures with Connections and “Doing Mathematics.” Descriptors of each level of the framework appear in Figure 1. Stein et al. (1996) argued that it was important to examine the cognitive demand required by tasks because of their influence on student learning:

> The mathematical tasks with which students become engaged determine not only what substance they learn but also how they come to think about, develop, use, and make sense of mathematics. Indeed, an important distinction that permeates research on academic tasks is the differences between tasks that engage students at a surface level and tasks that engage students at a deeper level by demanding interpretation, flexibility, the shepherding of resources, and the construction of meaning. (p. 459)

To date, the Mathematical Tasks Framework has not been used to analyze the levels of cognitive demand required by the tasks contained in a series of textbooks, let alone the probability tasks from series published over a 50-year period. Thus, in an effort to more fully describe the treatment of probability in textbooks, we made distinctions between those tasks that require students to (a) simply memorize information, (b) routinely perform algorithms without giving any attention to the meaning or development of the procedure, (c) focus on the meaning of a procedure or algorithm, and (d) explore and analyze the mathematical features of a situation.

4. METHODOLOGY

4.1. SAMPLE SELECTION

Recent Eras of Mathematics Education In order to determine historical trends in the treatment of probability in curricular materials, we selected two textbook series from each of the four most recent eras of mathematics education (Fey & Graeber, 2003; Payne, 2003): the New Math, Back to Basics, a focus on Problem Solving, and the advent of the National Council of Mathematics’ [NCTM] Standards.

The “New Math” era was so named by the contemporary popular media, as a descriptor of the innovative mathematics curricula that were being developed during this time period. Several of these curricula were developed as a response to the 1957 launch

Levels of Demands

Lower-level demands (Memorization):

- Involve either reproducing previously learned facts, rules, formulas, or definitions or committing facts, rules, formulas or definitions to memory.
- Cannot be solved using procedures because a procedure does not exist or because the time frame in which the task is being completed is too short to use a procedure.
- Are not ambiguous. Such tasks involve the exact reproduction of previously seen material, and what is to be reproduced is clearly and directly stated.
- Have no connection to the concepts or meaning that underlie the facts, rules, formulas, or definitions being learned or reproduced.

Lower-level demands (Procedures without Connections):

- Are algorithmic. Use of the procedure either is specifically called for or is evident from prior instruction, experience, or placement of the task.
- Require limited cognitive demand for successful completion. Little ambiguity exists about what needs to be done and how to do it.
- Have no connection to the concepts or meaning that underlie the procedure being used.
- Are focused on producing correct answers instead of on developing mathematical understanding.
- Require no explanations or explanations that focus solely on describing the procedure that was used.

Higher-level demands (Procedures with Connections):

- Focus students' attention on the use of procedures for the purpose of developing deeper levels of understanding of mathematical concepts and ideas.
- Suggest explicitly or implicitly pathways to follow that are broad general procedures that have close connections to underlying conceptual ideas as opposed to narrow algorithms that are opaque with respect to underlying concepts.
- Usually are represented in multiple ways, such as visual diagrams, manipulatives, symbols, and problem situations. Making connections among multiple representations helps develop meaning.
- Require some degree of cognitive effort. Although general procedures may be followed, they cannot be followed mindlessly. Students need to engage with conceptual ideas that underlie the procedures to complete the task successfully and that develop understanding.

Higher-level demands (Doing Mathematics):

- Require complex and nonalgorithmic thinking—a predictable, well-rehearsed approach or pathway is not explicitly suggested by the task, task instructions, or a worked-out example.
- Require students to explore and understand the nature of mathematical concepts, processes, or relationships.
- Demand self-monitoring or self-regulation of one's own cognitive processes.
- Require students to access relevant knowledge and experiences and make appropriate use of them in working through the task.
- Require students to analyze the task and actively examine task constraints that may limit possible solution strategies and solutions.
- Require considerable cognitive effort and may involve some level of anxiety for the student because of the unpredictable nature of the solution process required.

Smith and Stein (1998). Reprinted with permission from *Mathematics Teaching in the Middle School*, copyright 1998 by the National Council of Teachers of Mathematics. All rights reserved.

Figure 1. Characteristics of tasks at different levels of cognitive demand

of Sputnik and subsequent U.S. realization of the need for improvement in mathematics education (DeVault & Weaver, 1970; Osbourne & Crosswhite, 1970). Several facets of the New Math materials were met with intense opposition. A growing concern began to emerge from the public and elementary school teachers that students were unable to accurately compute (Payne, 2003). This growing concern blossomed into a full-fledged reactionary movement in the 1970s that focused students on the fundamentals of mathematics. For this reason, this era is referred to as "Back to Basics," where the basics were primarily defined as computational skills (Usiskin, 1985).

After a decade of focused attention on procedures and algorithms, the NCTM (1980) published *An Agenda for Action*, calling for a focus on problem solving in mathematics classes during the 1980s. Other organizations (College Board, 1983; National Academy of Sciences and National Academy of Engineering, 1982; National Commission on Excellence in Education, 1983; National Science Foundation and Department of Education, 1980) also issued reports and recommendations for mathematics education. Usiskin (1985) summarized these recommendations as follows: "Taken as a body, reports from inside and outside mathematics education agree almost unanimously that … emphasis should be shifted from rote manipulation to problem solving" (p. 15).

In 1989, the NCTM published *Curriculum and Evaluation Standards for School Mathematics*, calling for reform of mathematics education on a wide scale. In this document, the Council provided recommendations for mathematical content that ought to receive increased or decreased attention in the classroom and outlined important mathematical processes, such as problem solving and communication, that should be encouraged and fostered as students do mathematics. This document, along with *Professional Standards for Teaching Mathematics* (NCTM, 1991) and *Assessment Standards for School Mathematics* (NCTM, 1995) provided classroom teachers and mathematics educators with a conceptual anchor for reforming their practice. In an attempt to focus the reform of mathematics education into the new millennium, the NCTM (2000) published *Principles and Standards for School Mathematics*. This document represented further refinements of the earlier Standards documents in an integrated format, and provided more detailed narrative of the recommendations of the Council.

It is difficult to determine the precise beginning and end of these eras, and a significant event that marks the start of a new era (e.g., the publication of the *Curriculum and Evaluation Standards for School Mathematics* in 1989) does not necessarily immediately impact the textbooks that are published that year or the next. Nevertheless, we acknowledge the need to specify time frames for each era. Hereafter, we refer to the years 1957-1972 as the New Math era, 1973-1983 as the Back to Basics era, 1984-1993 as the Problem Solving era, and 1994-2004 as the Standards era. Table 1 displays the years that we designated as the terminal points of each era.

Table 1. Operational time frames for recent eras in mathematics education

Mathematics Education Era	Time Frame
New Math	1957-1972
Back to Basics	1973-1983
Problem Solving	1984-1993
Standards	1994-2004

For each era, we selected two series of mathematics textbooks: one series that was used by a relatively large proportion of middle-grade students in the United States, and one series that was different from "popular" textbooks at the time, possibly because of the authors' desire to reform mathematics education by providing alternative curricular materials. We refer to the former type as *popular*, and the latter as *alternative*. We examined both popular and alternative textbooks from each era in an attempt to gain a broad perspective on the treatment of probability topics for that era.

Popular Textbook Selection In this study, we define a popular textbook series as the mathematics textbook series having the largest market share during a given era. Hereafter, these textbooks are referred to as *popular.* Textbook market share data are available (Weiss, 1978, 1987; Weiss et al., 2001) and were used to determine which textbook series was the most popular during the Back to Basics, Problem Solving, and Standards eras. In the absence of market share data for the New Math era, the popular textbook series was determined by a "professional consensus" of mathematics educators familiar with the middle-grades curriculum during the past 50 years and affiliated with the Center for the Study of Mathematics Curriculum.

For each era, the popular textbooks that were considered for selection must be intended for use with students in grades 6, 7, and 8. Furthermore, these textbooks should have been written for the "average-level" student, that is, neither remedial nor accelerated. For this reason, algebra textbooks (i.e., textbooks that primarily focused on algebra, and are geared toward more mathematically advanced students in the middle grades) such as *Algebra 1/2* (Saxon, 1980) or *Algebra through Applications with Probability and Statistics* (Usiskin, 1979), for example, were not considered in this study.

Data from Weiss (1978, 1987) and Weiss et al. (2001) yielded the following sample of popular textbooks (see Table 2): *Holt School Mathematics* (Nichols et al., 1974a, 1974b, 1974c) for the Back to Basics era; *Mathematics Today* (Abbott and Wells, 1985a, 1985b, 1985c) published by Harcourt Brace Jovanovich for the Problem Solving era; and *Mathematics: Applications and Connections* (Collins et al., 1998a, 1998b, 1998c) for the Standards era. For the New Math era, a majority of those comprising the "professional consensus" stated that *Modern School Mathematics* (Dolciani, Beckenbach, Wooten, Chinn, & Markert, 1967a, 1967b; Duncan, Capps, Dolciani, Quast, & Zweng, 1967) was one of the most (if not *the* most) popular textbook series for middle-grades students during the New Math era.

In this study, we examined only the student editions of each textbook, because we were primarily interested in the tasks that students may have encountered as they used the textbooks. We did not examine the teacher's editions because students typically do not interact directly with the material within the teacher's edition; the teacher usually mediates this interaction. Although research indicates that teachers also mediate a student's interaction with the student's textbook edition by lowering the cognitive demand for tasks (e.g., Arbaugh, Lannin, Jones, & Park-Rogers, 2006; Stein & Smith, 1998; Stein et al., 2000), it would be impossible (in most cases) to document the myriad of interactions between teachers and curricular materials over the past several decades. For this reason, we focus solely on the student editions of the textbook and acknowledge that our study was not designed to capture any teacher actions regarding implementation of the curricula.

Alternative Textbook Selection As with the popular textbooks, the alternative series that were considered needed to be written for the "average-level" student in grades 6-8, and algebra textbooks were not considered. Additionally, we intended to examine textbooks that were part of a comprehensive mathematics series. Thus, we did not consider materials from the Middle Grades Mathematics Project (Phillips, Lappan, Winter, & Fitzgerald, 1986) or the Quantitative Literacy Series (Newman, Obremski, &

Scheaffer, 1987) because they were originally written as supplemental units, not as a comprehensive stand-alone curriculum.

Identifying textbook series that were “alternative” (i.e., series that were possibly innovative, influential, or offered as a departure from the popular series of the time) requires assigning a value judgment to that series. Such value judgments are subjective and vary among individuals. In order to counter the subjectivity of this process, the aforementioned “professional consensus” was solicited to identify alternative middle-grades mathematics textbook series for each of the eras of concern.

Results from the professional consensus yielded the following sets of alternative textbook series for each of the specified eras, as depicted in Table 2. *Mathematics for the Elementary School: Grade 6* (School Mathematics Study Group [SMSG], 1962) and *Mathematics for Junior High School* (SMSG, 1961a, 1961b) were created with support from the National Science Foundation (NSF) during the New Math era. The SMSG materials were used in many classrooms across the United States, and had substantial impact on the content of several commercially-developed textbooks (Payne, 2003). *Real Math* (Willoughby, Bereiter, Hilton, & Rubenstein, 1981, 1985a, 1985b) published by Open Court during the Back to Basics era, was offered as an alternative to popular textbooks which focused almost exclusively on computation, as stated in an advertisement from the October 1977 issue of *Arithmetic Teacher*. Saxon Publishers offered *Math 65, Math 76,* and *Math 87* (Hake & Saxon, 1985, 1987, 1991) during the Problem Solving era as alternative to the popular textbooks of the time, and focused on an incremental development of skills. *Connected Mathematics Project* materials (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998a, 1998b, 1998c, 1998d, 1998e, 1998f, 1998g, 1998h, 1998i, 1998j, 1998k, 1998l, 1998m, 1998n, 1998o, 1998p, 1998q, 1998r, 1998s, 1998t, 1998u, 1998v, 1998w, 1998x) were created with the support of the NSF during the Standards era, and had the largest market share of all such middle-grades mathematics materials. The *Connected Mathematics Project* units were divided into grade levels according to the authors’ suggested order in *Getting to Know Connected Mathematics* (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1996).

Table 2. Set of textbooks selected for analysis, with labels used for this study

Era	Type	Textbook Titles	Publisher
New Math	Popular	• *Modern School Mathematics: Structure and Use 6* • *Modern School Mathematics: Structure and Method 7 & 8*	Houghton Mifflin
	Alternative	• *Mathematics for the Elementary School, Grade 6* • *Mathematics for Junior High School, Vols. I & II*	Yale University Press
Back to Basics	Popular	• *Holt School Mathematics: Grades 6, 7, & 8*	Holt, Rinehart, & Winston
	Alternative	• *Real Math: Levels 6, 7, & 8*	Open Court
Problem Solving	Popular	• *Mathematics Today: Levels 6, 7, & 8*	Harcourt Brace Jovanovich
	Alternative	• *Math 65: An Incremental Development* • *Math 76: An Incremental Development* • *Math 87: An Incremental Development*	Saxon Publishers
Standards	Popular	• *Mathematics: Applications and Connections: Courses 1, 2, & 3*	Glencoe/ McGraw-Hill
	Alternative	• *Connected Mathematics*	Dale Seymour

4.2. ANALYSIS METHODS FOR IDENTIFICATION OF TASKS

Drawing heavily on the work of the QUASAR Project (e.g., Smith & Stein, 1998; Stein et al., 1996; Stein & Smith, 1998; Stein et al., 2000), we use the term *probability task* (or simply *task*) to refer to an activity, exercise, or set of exercises in a textbook that has been written with the intent of focusing a student's attention on a particular idea from probability. Any task that contained probability was considered a probability task, even if the main focus of the task was on another content area, such as geometry, combinatorics, or statistics. A probability task is not necessarily a single exercise in the textbook. A set of exercises that build on one another are considered as a single task. We have constructed such a task, as illustrated in Figure 2.

How likely is it that a chocolate chip will land on the flat side after being tossed in the air? Perform the following experiment and answer these questions to help formulate your answer to this question.

1. What are the possible outcomes for the landing position of a chocolate chip?
2. With your partner, toss 50 chocolate chips and record the landing position. How many chips landed on the flat side?
3. Based on your data, what is the experimental probability of a chocolate chip landing on the flat side?
4. As a class, pool your data. Based on the pooled data, what is the experimental probability of a chocolate chip landing on the flat side?
5. How does the experimental probability based on your data compare to the experimental probability based on the pooled data? How do you account for any differences?
6. Which of these experimental probabilities do you believe to be closest to the theoretical probability? Why? How could you obtain a better estimate of the theoretical probability?

Figure 2. Sample probability task

Likewise, a set of exercises that attend to the same topic but may be answered in isolation is considered as one task, as is the case in the task we constructed for Figure 3. Sections of probability lessons that contain narrative, such as definitions or written explanations of concepts and procedures, are not considered as probability tasks, although they are considered as portions of the textbook devoted to topics in probability.

A gumball machine contains four red gumballs, five blue gumballs, and six green gumballs. Rosalie selects one gumball from the machine at random.

1. What is the probability that the gumball is red?
2. What is the probability that the gumball is yellow?
3. What is the probability that the gumball is not blue?
4. What is the probability that the gumball is red or blue?

Figure 3. Sample probability task.

We examined each page of the selected textbooks for probability content. The portions of these textbooks that contained probability content were divided into discrete probability tasks by the first author and subsequently validated by the second author. As mentioned previously, these tasks may have consisted of several questions related to the same mathematical idea. Because of this distinction, in a given textbook the number of probability tasks that were identified was less than the number of questions, examples, and activities related to probability. Most probability tasks were located within lessons, in both the development (e.g., worked examples, activities) and assignment portion of lessons. Other probability tasks were not located in lessons, but in chapter reviews, assessments, and extension or enrichment activities.

4.3. CODING AND ANALYZING THE LEVEL OF COGNITIVE DEMAND OF PROBABILITY TASKS

We coded each task according to the level of cognitive demand that it required. According to the Levels of Demand criteria (Smith & Stein, 1998; see Figure 1), we indicated whether the task required Memorization (Low-M), Procedures without Connections (Low-P), Procedures with Connections (High-P), or "Doing Mathematics" (High-D). Tasks containing multiple questions were analyzed as a whole; therefore, we coded each task as requiring a single level of cognitive demand. The two researchers performed check-coding (Miles & Huberman, 1994) on tasks from two randomly selected textbooks from our sample by independently coding each task and then comparing assigned codes. Initial agreement was reached on the assignment of approximately 82% of the tasks, and 100% agreement was reached after discussion. The first author then proceeded in coding all probability tasks contained in the remaining textbooks in the sample.

5. RESULTS

5.1. NUMBER OF PROBABILITY TASKS IN EACH SERIES

Figure 4 displays the number of probability tasks for each series by grade level. Note that most series have the greatest number of probability tasks in the 8th grade textbook and the least in the 6th grade textbook, although this is not uniformly the case. The Standards-Alternative series has quite a different composition, with over half of the probability tasks located in the 7th grade textbook.

There were approximately equivalent numbers of probability tasks in the New Math-Popular, New Math-Alternative, Back to Basics-Popular, Back to Basics-Alternative, and Problem Solving-Popular textbook series. The Problem Solving-Alternative series had the fewest number of probability tasks (42) of all textbook series in the sample, less than half of the number in the New Math-Alternative and Back to Basics-Popular series, which ranked next to lowest in number of probability tasks. The number of probability tasks in five of the six textbooks from the Standards era was nearly equivalent to or greater than the number of tasks in an entire series from any of the other three eras. Furthermore, it should be noted that more than half of the probability tasks from the entire sample were located within textbooks from the Standards era.

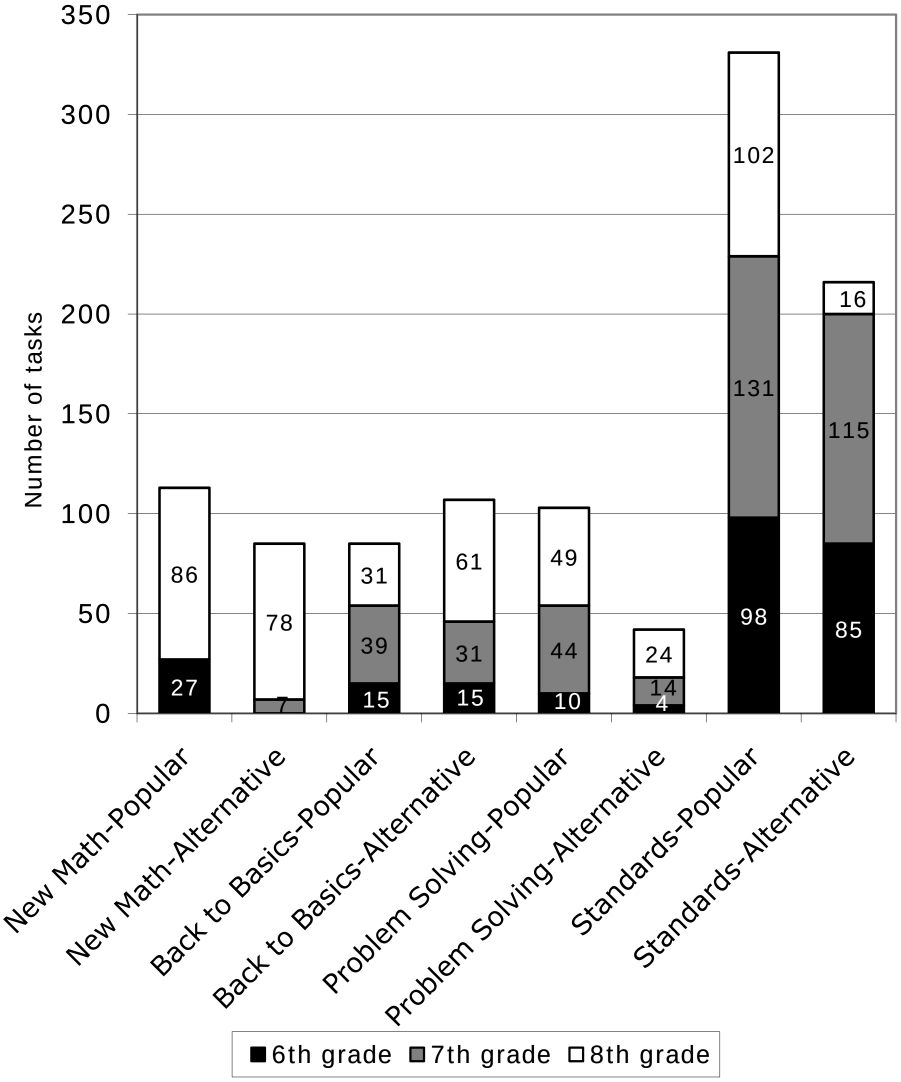


Figure 4. Number of probability tasks in each series, by grade level

5.2. DISTRIBUTION OF REQUIRED LEVELS OF COGNITIVE DEMAND WITHIN A TEXTBOOK SERIES

Within each era, and across the eras, the majority of tasks required low levels of cognitive demand, predominantly Procedures without Connections (Low-P). The Standards-Alternative series was an exception, with the majority of tasks requiring high levels of cognitive demand. The two tasks shown in Figure 5 both require lower levels of cognitive demand, Memorization (Low-M) and Procedures without Connections (Low-P). The rationale for this coding follows in the next few paragraphs.

The task in Question 3 was coded as Memorization because it was preceded by text that contained the definition of the term “dependent events,” a description of the procedure and a worked example incorporating the procedure for finding the probability of the occurrence of two dependent events. When working through the text sequentially, a student would first read the definition and worked example, and later read this task, prompting him or her to merely recall and provide the definition. This task could be completed by referring to the preceding text, or from memorizing the given procedure of "multiply the probability of the first event by the probability of the second event" (Collins et al., 1998c, p. 522).

Questions 10 through 14 were identified as a single task and coded as Procedures without Connections. This task is algorithmic, and there is little ambiguity on how to complete the task. As described above, the text that precedes this task includes a description and worked example of the exact procedure that is to be followed. Furthermore, this task requires no explanations and appears to focus more on correct answers than fostering the development of mathematical understanding.

3. Tell how to find the probability of two dependent events.

In a bag there are 5 red marbles, 2 yellow marbles, and 1 blue marble. Once a marble is selected, it is not replaced. Find the probability of each outcome.

10. a red marble and then a yellow marble
11. a blue marble and then a yellow marble
12. a red marble and then a blue marble
13. any color marble except yellow and then a yellow marble
14. a red marble three times in a row

From *Mathematics: Applications and Connections, Course 3* © 1998, Collins, Dristas, Frey-Mason, Howard, McClain, Molina, et al. Published by Glencoe/McGraw-Hill. Used by permission.

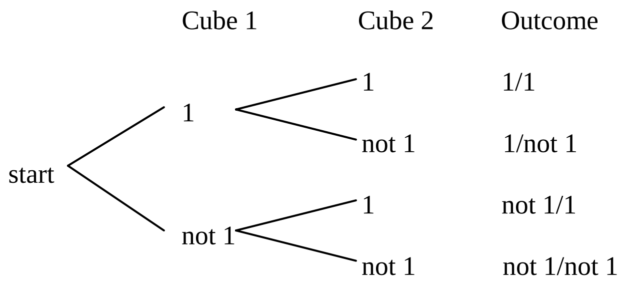
Figure 5. Examples of tasks that require low levels of cognitive demand

As stated previously, the majority of tasks in seven of the eight textbook series required low levels of cognitive demand, primarily Procedures without Connections. In contrast to these series, most probability tasks in the Standards-Alternative series required higher levels of cognitive demand. Examples of two tasks from this series requiring higher levels of cognitive demand are shown in Figure 6. The rationale behind this coding follows below.

The task in Question 19 was coded as Procedures with Connections (High-P) because it addresses a common misconception (all outcomes are equally likely) without suggesting a pathway to the solution, either in the task itself or on preceding pages. This task utilizes multiple representations of the problem situation (a counting tree and an organized listing of the complete sample space), and requires some degree of cognitive effort, as no algorithm or procedure has been previously given that addresses this situation in this form.

The task in Problem 6.1 and Problem 6.1 Follow-Up, coded as "Doing Mathematics" (High-D), requires complex and nonalgorithmic thinking. Prior to this task, the textbook authors have provided tasks to allow students to conduct experiments and calculate and compare experimental and theoretical probabilities, but this is the first request to *create* a simulation. Students are required to utilize their prior knowledge and apply it to this task. Furthermore, this task requires that students understand several mathematical concepts, including the set of possible outcomes of this game, the expected number of wins in 100 trials, and the fairness of the game. Finally, students are instructed to provide justifications for their reasoning in questions 2 and 3.

19. Tricia wants to determine the probability of getting two 1s when two number cubes are rolled. She made a counting tree and used it to list the possible outcomes.



She says that, since there are four possible outcomes, the probability of getting 1 on both number cubes is $\frac{1}{4}$. Is Tricia right? Why or why not?

- -

Tawanda's Toys is having a contest! Any customer who spends at least $10 receives a scratch-off game card. Each card has five gold spots that reveal the names of video games when they are scratched. Exactly two spots match on each card. A customer may scratch off only two spots on a card; if the spots match, the customer wins the video game under those spots.

Problem 6.1
If you play this game once, what is your probability of winning? To answer this question, do the following two things:

A. Create a way to simulate Tawanda's contest, and find the experimental probability of winning.
B. Analyze the different ways you can scratch off two spots, and find the theoretical probability of winning a prize with one game card.

Problem 6.1 Follow-Up

1. a. If you play Tawanda's scratch-off game 100 times, how many video games would you expect to win?
   b. How much money would you have to spend to play the game 100 times?
2. Tawanda wants to be sure she will not lose money on her contest. The video games she gives as prizes cost her about $15 each. Will Tawanda lose money on this contest? Why or why not?
3. Suppose you play Tawanda's game 20 times and never win. Would you conclude that the game is unfair? For example, would you think that there were not two matching spots on every card? Why or why not?

From *Connected Mathematics: What Do You Expect? Probability and Expected Value* © 1998 by Michigan State University, Lappan, Fey, Fitzgerald, Friel, and Phillips. Published by Pearson Education, Inc., publishing as Pearson Prentice Hall. Used by permission.

Figure 6. Examples of tasks that require high levels of cognitive demand

Table 3 displays the percentage of tasks coded at each level of cognitive demand for each series. Using the Mann-Whitney *U* test (Hinkle, Wiersma, & Jurs, 1988; also named the "Mann, Whitney, and Wilcoxon test" in Hogg & Tanis, 1993), we determined that the distributions of required levels of cognitive demand were not significantly different for the three textbooks within a given series, with $p > 0.18$ in each case. Accordingly, the data presented in Table 3 represent all tasks in the series, without disaggregating by grade level.

Table 3. Percentage of tasks coded at each level of cognitive demand

	New Math		Back to Basics		Problem Solving		Standards	
	Pop.	Alt.	Pop.	Alt.	Pop.	Alt.	Pop.	Alt.
High-D	0	0	0	4	0	0	2	12
High-P	6	14	0	22	2	0	15	47
Low-P	83	82	95	74	97	98	75	40
Low-M	11	4	5	0	1	2	8	1

Note that the Back to Basics-Alternative series did not contain any tasks coded at the Memorization level (Low-M), and three series contained tasks coded at the highest level–"Doing Mathematics" (High-D). Furthermore, the Back to Basics-Popular and Problem Solving-Alternative series contained no probability tasks that required high levels of cognitive demand.

The composition of tasks found at each level of cognitive demand was very similar for the two series in the Problem Solving era; the New Math era series were also similar in this composition. In terms of cognitive demand, the series within both the Back to Basics and Standards eras appeared to be quite different, with the alternative series tending to have greater proportions of tasks that require higher levels of cognitive demand than the contemporary popular series.

5.3. TRENDS IN REQUIRED LEVELS OF COGNITIVE DEMAND OVER TIME

Typically, the most common level of cognitive demand required by probability tasks was Procedures without Connections (Low-P). The Standards-Alternative series was an exception, with nearly half of all tasks coded as Procedures with Connections (High-P). The majority of tasks that required high levels of cognitive demand were located within the series of the Standards era, as were the majority of tasks that were analyzed for this study. More specifically, there were a greater number of tasks, but not necessarily a greater percentage of tasks, that required higher levels of cognitive demand in the two Standards-era textbook series. The Standards-Popular series is a case of this phenomenon. It contains more tasks requiring higher levels of cognitive demand than any series from a previous era, but a smaller percentage of higher level tasks than the Back to Basics-Alternative series.

6. DISCUSSION

6.1. INTERPRETATION OF INCREASED NUMBER OF TASKS IN MORE RECENT SERIES

There was a dramatic increase in number of probability tasks in the textbooks from the Standards era, compared to the three previous eras of mathematics education. In particular, over half of all of the tasks analyzed in this study were located within textbooks from the Standards era. This increase in attention to probability appears to have coincided with the release of national recommendations such as NCTM (1989) that advocated the inclusion of probability in the middle grades. Although the design of this study did not allow for the identification of causal factors, the proliferation of probability tasks within the Standards era appears to be consistent with the contemporary recommendations for the inclusion of probability topics within the middle grades mathematics curriculum.

6.2. INTERPRETATION OF STABILITY OF DISTRIBUTION OF LEVEL OF COGNITIVE DEMAND WITHIN EACH SERIES, BUT DIFFERENCES BETWEEN SERIES

As stated above, across the three grade levels of each series, the distributions of required levels of cognitive demand of probability tasks were not significantly different. For most series, probability tasks required predominantly low levels. In the Standards-Alternative series, however, a majority of probability tasks (59%) required high levels of cognitive demand. Therefore, the Standards-Alternative series adhered to the recommendations of Stein et al. (2000) that students at each grade level should have opportunities to "engage with tasks that lead to deeper, more generative understandings regarding the nature of mathematical processes, concepts, and relationships" (p. 15). Furthermore, the use of tasks that require higher levels of cognitive demand in instruction supports the development of conceptual understanding that is called for by the NCTM (1989, 2000).

It is likely not a coincidence that the series with the highest distribution of required levels of cognitive demand (Standards-Alternative) was the same series that received the highest quality ratings in the Project 2061 study of mathematics textbooks for middle grades students (AAAS, 2000). The Project 2061 study did not examine the treatment of probability, but instead focused on number, algebra, and geometry. Although numerous criteria were used to render quality ratings, the results from our study indicate that the probability portions of this series may be of similar high quality. Additionally, Project 2061 researchers found that two other series included in our study (Standards-Popular and a revised edition of Problem Solving-Alternative) were of lower overall quality than the Standards-Alternative series; their findings coincide with results from our study that the probability tasks contained in the Problem Solving-Alternative and Standards-Popular series required significantly lower levels of cognitive demand than the Standards-Alternative series. Although the distribution of required levels of cognitive demand is not equivalent to the quality of instruction in a textbook, these measures are similar in that they address the potential opportunities for students to develop deeper understandings of mathematical content.

In the Back to Basics and Standards eras, there were significant differences in the distribution of required levels of cognitive demand for probability tasks between the popular and alternative series ($U = 3199.5$, $Z = -5.443$, $p < 0.001$ and $U = 19661.5$, $Z = -10.273$, $p < 0.001$, respectively). In each case, the alternative series had a higher distribution of required levels of cognitive demand. This may reflect the desires of the authors to offer something truly different. These alternative series presented more than a new sequence of topics or additional topics; indeed, the *nature* of the tasks within these textbook series was substantially different. This lends credence to the notion that these series represented true alternatives to the contemporary popular series.

7. RECOMMENDATIONS AND IMPLICATIONS

With the exception of the Standards-Alternative series, the vast majority of tasks in each series were characterized as requiring low levels of cognitive demand, usually at the Procedures without Connections level. Indeed, *all* tasks in the Back to Basics-Popular and Problem Solving-Alternative series were coded as Procedures without Connections, save five tasks within the Back to Basics-Popular series and one task in the Problem Solving-Alternative series coded as Memorization. Stein, Grover, and Henningsen (1996) found that the level of cognitive demand of a task as written tends to either stay the same or decline when implemented by the teacher. For this reason, textbooks should include tasks that require high levels of cognitive demand, and thus provide potential opportunities for students to experience mathematics as more than a set of unrelated procedures and facts. The inclusion of tasks that require high levels of cognitive demand has the potential to foster a more connected view of mathematics as related, meaningful concepts and procedures useful for solving many types of problems.

Heretofore there is no research documenting the impact of specific curricular tasks on student learning in probability, delineating between the kinds of reasoning tasks at each level may promote. Nevertheless, it seems reasonable to conjecture that the nature of a set of probability tasks might influence students' views of probability, and promote a more classical, frequentist, or subjective approach (for a more detailed description, see Batanero, Henry, & Parzysz, 2005; Borovcnik, Bentz, & Kapadia, 1991). For example, low-level tasks, including Procedures without Connections, might promote a classical, deterministic outlook, in which students rely on calculations of theoretical probabilities with little or no appreciation for how much variability might be expected in repeated trials of a probability experiment. On the other hand, higher-level tasks, such as Doing Mathematics, might foster a frequentist view in which students grapple with disparities between empirically-derived and theoretically-derived probabilities. This latter approach not only aligns with recent curriculum frameworks (e.g., NCTM, 2000), it is also consistent with research on learning probability that advocates teachers (a) "make connections between probability and statistics," (b) "introduce probability through data," and (c) "adapt a problem-solving approach to probability" (Shaughnessy, 2003, p. 224). Indeed, there is a growing body of evidence (e.g., delMas & Bart, 1989; Pratt, 2000; Pratt, 2005; Stohl & Tarr, 2002; Yáñez, 2002) to support the role of simulations as a means of fostering more sophisticated understanding of probability concepts in place of the well-documented *equiprobability bias* (Lecoutre, 1992), *outcome approach* (Konold, 1991), and *representativeness* (e.g., Konold et al., 1993). Therefore, we argue that more recent curricular materials containing higher-level probability tasks have the potential to promote sound probabilistic reasoning, challenging some of the primary intuitions students bring to the classroom, and preparing them to make sense of the chance variation and random phenomena they will inevitably face in the real world.

Results of this study revealed differences in the extent and nature of the treatment of probability across series. Prospective and practicing mathematics teachers may benefit from applying portions of the framework used in this study to analyze curricula and, in doing so, dispel the notion that all textbooks are created alike. In particular, determining the levels of cognitive demand required by tasks revealed that for most series, the majority of probability tasks required low levels of cognitive demand. By conducting similar analyses, it is possible prospective mathematics teachers will be more prepared to scrutinize their own textbook and realize the need to increase the levels of cognitive demand of tasks. Such teacher education activities form an important part of the call from researchers for prospective teachers to critically analyze textbooks and other curriculum materials (Lloyd & Behm, 2005; Remillard, 2004), In addition, preservice teachers with experience in analyzing curricula may be better prepared to assist in the selection of curriculum in their future positions, and less likely to examine only surface characteristics of textbooks.

Only two textbook series were selected from each era for the sample of this study. The mathematics curriculum of a particular era may be more fully described if more series, particularly more popular series, from each era were examined. Furthermore, the "alternative" curriculum may be more appropriately characterized by analyzing commonly used supplemental materials that did not fit the selection criteria to be included within this sample. For example, several mathematics educators familiar with middle grades mathematics curricula identified the *Middle Grades Mathematics Project* (Phillips et al., 1986) and *Transition Mathematics* (Usiskin et al., 1995) as widely used alternative curricula, but neither set of materials was written as a comprehensive curriculum for students in grades 6, 7, and 8. Thus, research should analyze a broader range of curricular materials, both popular and alternative, in order to more fully characterize the mathematics curriculum of particular eras of mathematics education.

For most of the series, no significant difference was found between the distribution of required levels of cognitive demand of tasks in the development and assignment portions of lessons. In each of the textbook series in the first three eras and the Standards-Popular series, $p > 0.05$. In the Standards-Alternative series ($U = 2821$, $Z = -3.03$, $p < 0.01$), the tasks in the development portions of lessons tended to require higher levels of cognitive demand than tasks within the assignment portions. Further research should examine the distribution of required levels of cognitive demand for assessment tasks as well, and compare these distributions to those of tasks within the development and assignment portions of lessons. Results from this study revealed that relatively few probability tasks were written for the purpose of summative assessment. For this reason, future research may need to examine ancillary materials for assessment items. Such analyses may reveal a potential mismatch between assessment and instruction. In particular, it may be that tasks within lessons require high levels of cognitive demand, whereas assessment tasks require low levels of cognitive demand.

Finally, this study focused on describing the intended probability curriculum as present in middle grades mathematics textbooks. Informed by the results of this study, future research should investigate the enacted probability curriculum as presented by teachers in contemporary classrooms. This enacted curriculum should include more than the particular chapters and lessons that were covered, but also which tasks were assigned, and why teachers chose to include or omit particular portions of the intended probability curriculum from their instruction. Such an analysis of the enacted curriculum was beyond the scope of this study, but it needs to be examined in order to more precisely determine students' opportunity to learn.

8. CONCLUSION

This study addressed an existing void in the research base by analyzing the treatment of probability within middle grades mathematics textbooks from a historical perspective. Moreover, it represented the first documented attempt to analyze the levels of cognitive demand required by probability tasks within textbooks published across four recent eras of mathematics education. Information about the levels of cognitive demand required by tasks within a textbook may prove to be one measure of the quality of the mathematics presented within a textbook or textbook series. Indeed, a significant result of our study was the increase in the number and proportion of tasks that required high levels of cognitive demand within the alternative textbook series from the Back to Basics and Standards eras. In essence, these two series may be viewed as models for the development of future curricula because they challenge students to move beyond the development of mere procedural knowledge of probability.

In this era of *Standards*, middle grades mathematics textbooks are devoting more attention to probability and requiring high levels of cognitive demand. Whether attributable to the recommendations of researchers or professional organizations, the alternative textbook series of the Back to Basics and Standards eras in this study demonstrate a markedly different approach to the level of cognitive demand required by probability tasks, and the alternative series from the Standards era provided many more opportunities for students to engage in these types of tasks. What remains to be documented is the impact of such curricular materials on student learning, particularly in reference to supporting students' understanding of probability concepts, so that probability misconceptions will become less prevalent among people of all ages.

ACKNOWLEDGEMENTS

The development of this paper was made possible with support from the Center for the Study of Mathematics Curriculum (NSF Award No. ESI-0333879). The views and conclusions expressed are those of the authors and not necessarily those of the National Science Foundation. The mathematics educators who provided their insight for the textbook sample selection were as follows: Glenda Lappan, Betty Phillips, and Sharon Senk (Michigan State University); Zalman Usiskin (University of Chicago); Denisse Thompson (University of South Florida); Douglas Grouws, Barbara Reys, and Robert Reys (University of Missouri); and Christian Hirsch (Western Michigan University).

REFERENCES

Abbott, J. S., & Wells, D. W. (1985a). *Mathematics today: Level 6.* Orlando, FL: Harcourt Brace Jovanovich.

Abbott, J. S., & Wells, D. W. (1985b). *Mathematics today: Level 7.* Orlando, FL: Harcourt Brace Jovanovich.

Abbott, J. S., & Wells, D. W. (1985c). *Mathematics today: Level 8.* Orlando, FL: Harcourt Brace Jovanovich.

Adams, L., Tung, K. K., Warfield, V. M., Knaub, K., Mudavanhu, B., & Yong, D. (2000). *Middle school mathematics comparisons for Singapore Mathematics, Connected Mathematics Program, and Mathematics in Context (including comparisons with the NCTM Principles and Standards 2000). A report to NSF, November 2, 2000.* Unpublished manuscript, Seattle, WA.

American Association for the Advancement of Science: Project 2061. (2000). *Middle grades mathematics textbooks: A benchmarks-based evaluation.* Washington, DC: Author.

Arbaugh, F., Lannin, J., Jones, D. L., & Park-Rogers, M. (2006). Examining instructional practices in Core-Plus lessons: Implications for professional development. *Journal of Mathematics Teacher Education, 9*(6), 517-550.

Batanero, C., Henry, M., & Parzysz, B. (2005).The nature of chance and probability. In G. A. Jones (Ed.) *Exploring probability in school: Challenges for teaching and learning,* (pp. 15-37). Netherlands: Kluwer Academic Publishers.

Borovcnik, M., Bentz, H. J., & Kapadia, R. (1991). A probabilistic perspective. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: Probability in education* (pp. 27-71). Boston: Kluwer Academic Publishers.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B., & Johnson, M. (2001). *The nation's report card: Mathematics 2000.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Jr., Lindquist, M. M., & Reys, R. E. (1981). *Results from the second mathematics assesment of the National Assessment of Educational Progress*. Reston, VA: National Council of Teachers of Mathematics.

Clopton, P., McKeown, E., McKeown, M., & Clopton, J. (1999a). *Mathematically correct fifth grade mathematics review.* [Online: http://mathematicallycorrect.com/books5.htm]

Clopton, P., McKeown, E., McKeown, M., & Clopton, J. (1999b). *Mathematically correct second grade mathematics review.* [Online: http://mathematicallycorrect.com/books2.htm]

Clopton, P., McKeown, E., McKeown, M., & Clopton, J. (1999c). *Mathematically correct seventh grade mathematics review.* [Online: http://mathematicallycorrect.com/books7.htm]

Collins, W., Dristas, L., Frey-Mason, P., Howard, A. C., McClain, K., Molina, D. D., et al. (1998a). *Mathematics: Applications and connections, course 1.* New York: Glencoe/McGraw Hill.

Collins, W., Dristas, L., Frey-Mason, P., Howard, A. C., McClain, K., Molina, D. D., et al. (1998b). *Mathematics: Applications and connections, course 2.* New York: Glencoe/McGraw Hill.

Collins, W., Dristas, L., Frey-Mason, P., Howard, A. C., McClain, K., Molina, D. D., et al. (1998c). *Mathematics: Applications and connections, course 3.* New York: Glencoe/McGraw Hill.

Conference Board of the Mathematical Sciences. (2001). *The mathematical education of teachers.* Washington, DC: Mathematical Association of America.

delMas, R. C., & Bart, W. M. (1989). The role of an evaluation exercise in the resolution of misconceptions of probability. *Focus on Learning Problems in Mathematics, 11,* 39-53.

DeVault, M. V., & Weaver, J. F. (1970). Forces and issues related to curriculum and instruction, K-6. In P. S. Jones & A. F. Coxford, Jr. (Eds.), *A history of mathematics education in the United States and Canada: Thirty-second yearbook* (pp. 90-152). Washington, DC: National Council of Teachers of Mathematics.

Dolciani, M. P., Beckenbach, E. F., Wooten, W., Chinn, W. G., & Markert, W. (1967a). *Modern school mathematics: Structure and method 7.* Boston: Houghton Mifflin.

Dolciani, M. P., Beckenbach, E. F., Wooten, W., Chinn, W. G., & Markert, W. (1967b). *Modern school mathematics: Structure and method 8.* Boston: Houghton Mifflin.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*(2), 159-199.

Doyle, W. (1988). Work in mathematics classes: The context of students' thinking during instruction. *Educational Psychologist, 23*(2), 167-180.

Duncan, E. R., Capps, L. R., Dolciani, M. P., Quast, W. G., & Zweng, M. J. (1967). *Modern school mathematics: Structure and use 6.* Boston: Houghton Mifflin.

Fey, J. T., & Graeber, A. O. (2003). From the new math to the *Agenda for Action*. In G. M. A. Stanic & J. Kilpatrick (Eds.), *A history of school mathematics* (Vol. 1, pp. 521-558). Reston, VA: National Council of Teachers of Mathematics.

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.

Grouws, D. A., & Smith, M. S. (2000). Findings from NAEP on the preparation and practices of mathematics teachers. In E. A. Silver & P. A. Kenney (Eds.), *Results from the Seventh Mathematics Assessment of the National Assessment of Education Progress* (pp. 107-141). Reston, VA: National Council of Teachers of Mathematics.

Hake, S., & Saxon, J. (1985). *Math 76: An incremental development*. Norman, OK: Saxon Publishers, Inc.

Hake, S., & Saxon, J. (1987). *Math 65: An incremental development*. Norman, OK: Saxon Publishers, Inc.

Hake, S., & Saxon, J. (1991). *Math 87: An incremental development*. Norman, OK: Saxon Publishers, Inc.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1988). *Applied statistics for the behavioral sciences* (2nd ed.). Boston: Houghton Mifflin Company.

Hogg, R. V., & Tanis, E. A. (1993). *Probability and statistical inference* (4th ed.). New York: Macmillan Publishing Company.

Konold, C. (1983). Conceptions about probability: Reality between a rock and a hard place. (Doctoral dissertation, University of Massachusetts, 1983). *Dissertation Abstracts International*, 43, 4179b.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*, 59-98.

Konold, C. (1991). Understanding students' beliefs about probability. *Cognition and Instruction*, *6*, 59-98.

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

Lappan, G., Fey, J., Fitzgerald, W. M., Friel, S. N., & Phillips, E. (1996). *Getting to know Connected Mathematics: A guide to the Connected Mathematics curriculum.* Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998a). *Accentuate the negative*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998b). *Bits and pieces I.* Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998c). *Bits and pieces II*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998d). *Clever counting*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998e). *Comparing and scaling*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998f). *Covering and surrounding*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998g). *Data about us*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998h). *Data around us*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998i). *Filling and wrapping*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998j). *Frogs, fleas, and painted cubes*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998k). *Growing, growing, growing*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998l). *How likely is it?* Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998m). *Hubcaps, kaleidoscopes, and mirrors*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998n). *Looking for Pythagoras*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998o). *Moving straight ahead*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998p). *Prime time*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998q). *Ruins of Montarek*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998r). *Samples and populations*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998s). *Say it with symbols*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998t). *Shapes and designs*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998u). *Stretching and shrinking*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998v). *Thinking with mathematical models*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998w). *Variables and patterns*. Palo Alto, CA: Dale Seymour Publications.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998x). *What do you expect?* Palo Alto, CA: Dale Seymour Publications.

Lecoutre, M. (1992). Cognitive models and problem spaces in “purely random” situations. *Educational Studies in Mathematics, 23*, 557-568.

Lloyd, G. M. (1999). Two teachers’ conceptions of a reform-oriented curriculum: Implications for mathematics teacher development. *Journal of Mathematics Teacher Education, 2,* 227-252.

Lloyd, G. M., & Behm, S. L. (2005). Preservice elementary teachers’ analysis of mathematics instructional materials. *Action in Teacher Education, 26*(4), 48-62.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.

National Council of Teachers of Mathematics. (1980). *An agenda for action: Recommendations for school mathematics of the 1980s*. Reston, VA: Author.

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

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.

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

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academy Press.

Newman, C. M., Obremski, T. E., & Scheaffer, R. L. (1987). *Exploring probability*. Palo Alto, CA: Dale Seymour.

Nichols, E. D., Anderson, P. A., Dwight, L. A., Flourney, F., Kalin, R., Schluep, J., et al. (1974a). *Holt school mathematics: Grade 6*. New York: Holt, Rinehart and Winston.

Nichols, E. D., Anderson, P. A., Dwight, L. A., Flourney, F., Kalin, R., Schluep, J., et al. (1974b). *Holt school mathematics: Grade 7*. New York: Holt, Rinehart and Winston.

Nichols, E. D., Anderson, P. A., Dwight, L. A., Flourney, F., Kalin, R., Schluep, J., et al. (1974c). *Holt school mathematics: Grade 8*. New York: Holt, Rinehart and Winston.

Osborne, A. R., & Crosswhite, F. J. (1970). Forces and issues related to curriculum and instruction, 7-12. In P. S. Jones & A. F. Coxford, Jr. (Eds.), *A history of mathematics* *education in the United States and Canada: Thirty-second yearbook* (pp. 153-297). Washington, DC: National Council of Teachers of Mathematics.

Payne, J. N. (2003). The new math and its aftermath, grades K-8. In G. M. A. Stanic & J. Kilpatrick (Eds.), *A history of school mathematics* (Vol. 1, pp. 559-598). Reston, VA: National Council of Teachers of Mathematics.

Phillips, E., Lappan, G., Winter, M. J., & Fitzgerald, W. (1986). *Probability*. Menlo Park, CA: Addison-Wesley.

Porter, A. C. (2006). Curriculum assessment. In J. Green, G. Camilli & P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141-159). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pratt, D. (2000). Making sense of the total of two dice. *Journal of Research in Mathematics Education, 31*, 602-625.

Pratt, D. (2005). How do teachers foster students' understanding of probability? In G. A. Jones (Ed.) *Exploring probability in school: Challenges for teaching and learning,* (pp. 171-189). Netherlands: Kluwer Academic Publishers.

Remillard, J. T., & Bryans, M. B. (2004). Teachers' orientations toward mathematics curriculum materials: Implications for teacher learning. *Journal for Research in Mathematics Education, 35*, 352-388.

Robinson, E., & Robinson, M. (1996). *A guide to standards-based instructional materials in secondary mathematics*. Unpublished manuscript.

Robitaille, D. F., & Travers, K. J. (1992). International studies of achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 687-723). Reston, VA: National Council of Teachers of Mathematics.

Saxon, J. (1980). *Algebra 1/2*. Norman, OK: Saxon Publishers, Inc.

School Mathematics Study Group. (1961a). *Mathematics for junior high school* (Vol. I). New Haven, CT: Yale University Press.

School Mathematics Study Group. (1961b). *Mathematics for junior high school* (Vol. II). New Haven, CT: Yale University Press.

School Mathematics Study Group. (1962). *Mathematics for the elementary school: Grade 6*. New Haven, CT: Yale University Press.

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

Shaughnessy, J. M. (2003). Research on students' understandings of probability. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 216-226). Reston, VA: National Council of Teachers of Mathematics.

Smith, M. S., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School, 3*, 344-350.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488.

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection. *Mathematics Teaching in the Middle School, 3*, 268-275.

Stein, M. K., Smith, M. S., Henningsen, M., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development.* New York: Teachers College Press.

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior 21*, 319-337.

Tyson-Bernstein, H., & Woodward, A. (1991). Ninteenth century policies for twenty-first century practice: The textbook reform dilemma. In P. G. Altbach, G. P. Kelly, H. G. Petrie & L. Weis (Eds.), *Textbooks in American society: Politics, policy, and pedagogy*. Albany, NY: State University of New York Press.

Usiskin, Z. (1979). *Algebra through applications with probability and statistics*. Reston, VA: National Council of Teachers of Mathematics.

Usiskin, Z. (1985). We need another revolution in secondary school mathematics. In C. R. Hirsch & M. J. Zweng (Eds.), *The secondary school mathematics curriculum: 1985 yearbook of the National Council of Teachers of Mathematics* (pp. 1-21). Reston, VA: National Council of Teachers of Mathematics.

Usiskin, Z., Feldman, C. H., Flanders, J., Polonsky, L., Porter, S., & Viktora, S. S. (1995). *Transition mathematics*. Glenview, IL: Scott, Foresman.

Valverde, G. A., Bianchi, L. J., Wolfe, R. G., Schmidt, W. H., & Houang, R. T. (2002). *According to the book: Using TIMSS to investigate the translation of policy into practice through the world of textbooks*. Boston: Kluwer Academic Publishers.

Weiss, I. R. (1978). *Report of the 1977 National Survey of Science, Mathematics, and Social Studies Education*. Research Triangle Park, NC: Center for Educational Research and Evaluation.

Weiss, I. R. (1987). *Report of the 1985-1986 National Survey of Science and Mathematics Education*. Research Triangle Park, NC: Research Triangle Institute.

Weiss, I. R., Banilower, E. R., McMahon, K. C., & Smith, P. S. (2001). *Report of the 2000 national survey of science and mathematics education*. Chapel Hill, NC: Horizon Research, Inc.

Willoughby, D., Bereiter, C., Hilton, P., & Rubenstein, J. H. (1981). *Real math: Level 6*. La Salle, IL: Open Court.

Willoughby, D., Bereiter, C., Hilton, P., & Rubenstein, J. H. (1985a). *Real math: Level 7*. La Salle, IL: Open Court.

Willoughby, D., Bereiter, C., Hilton, P., & Rubenstein, J. H. (1985b). *Real math: Level 8*. La Salle, IL: Open Court.

Yáñez, G. C. (2002). Some challenges for the use of computer simulations for solving conditional probability problems. In D. S. Mewborn et al. (Eds.), *Proceedings of the 24th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Educati*on (pp. 1255-1266). Athens, GA: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

DUSTIN L. JONES
Sam Houston State University
Mathematics & Statistics
Box 2206
Huntsville, TX 77341
USA

ASSESSING STUDENTS' CONCEPTUAL UNDERSTANDING AFTER A FIRST COURSE IN STATISTICS

ROBERT DELMAS
University of Minnesota
delma001@umn.edu

JOAN GARFIELD
University of Minnesota
jbg@umn.edu

ANN OOMS
Kingston University
a.ooms@kingston.ac.uk

BETH CHANCE
California Polytechnic State University
bchance@calpoly.edu

ABSTRACT

This paper describes the development of the CAOS test, designed to measure students' conceptual understanding of important statistical ideas, across three years of revision and testing, content validation, and realiability analysis. Results are reported from a large scale class testing and item responses are compared from pretest to posttest in order to learn more about areas in which students demonstrated improved performance from beginning to end of the course, as well as areas that showed no improvement or decreased performance. Items that showed an increase in students' misconceptions about particular statistical concepts were also examined. The paper concludes with a discussion of implications for students' understanding of different statistical topics, followed by suggestions for further research.

Keywords: *Statistics education research; Assessment; Conceptual understanding; Online test*

1. INTRODUCTION

What do students know at the end of a first course in statistics? How well do they understand the important concepts and use basic statistical literacy to read and critique information in the world around them? Students' difficulty with understanding probability and reasoning about chance events is well documented (Garfield, 2003; Konold, 1989, 1995; Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993; Pollatsek, Konold, Well, & Lima, 1984; Shaughnessy, 1977, 1992). Studies indicate that students also have difficulty with reasoning about distributions and graphical representations of distributions (e.g., Bakker & Gravemeijer, 2004; Biehler, 1997; Ben-Zvi 2004; Hammerman & Rubin, 2004; Konold & Higgins, 2003; McClain, Cobb, & Gravemeijer, 2000), and understanding concepts related to statistical variation such as measures of variability (delMas & Liu, 2005; Mathews & Clark, 1997; Shaughnessy, 1977), sampling variation (Reading & Shaughnessy, 2004; Shaughnessy, Watson, Moritz, & Reading, 1999), and sampling distributions (delMas, Garfield, & Chance, 1999; Rubin, Bruce, & Tenney, 1990; Saldanha & Thompson, 2001). There is evidence that instruction can have positive effects on students' understanding of these concepts (e.g., delMas & Bart, 1989; Lindman & Edwards, 1961; Meletiou-Mavrotheris & Lee, 2002; Sedlmeier, 1999), but many students can still have conceptual difficulties even after the use of innovative instructional approaches and software (Chance, delMas, & Garfield, 2004; Hodgson, 1996; Saldanha & Thompson, 2001).

Partially in response to the difficulties students have with learning and understanding statistics, a reform movement was initiated in the early 1990s to transform the teaching of statistics at the introductory level (e.g., Cobb, 1992; Hogg, 1992). Moore (1997) described the reform movement as primarily having made changes in content, pedagogy, and technology. As a result, Scheaffer (1997) observed that there is more agreement today among statisticians about the content of the introductory course than in the past. Garfield (2001), in a study conducted to evaluate the effect of the reform movement, found that many statistics instructors are aligning their courses with reform recommendations regarding technology, and, to some extent, with teaching methods and assessment. Although there is evidence of changes in statistics instruction, a large national study has not been conducted on whether these changes have had a positive effect on students' statistical understanding, especially with difficult concepts like those mentioned above.

One reason for the absence of research on the effect of the statistics reform movement may be the lack of a standard assessment instrument. Such an instrument would need to measure generally agreed upon content and learning outcomes, and be easily administered in a variety of institutional and classroom settings. Many assessment instruments have consisted of teachers' final exams that are often not appropriate if they focus on procedures, definitions, and skills, rather than conceptual understanding (Garfield & Chance, 2000). The Statistical Reasoning Assessment (SRA) was one attempt to develop and validate a measure of statistical reasoning, but it focuses heavily on probability, and lacks items related to data production, data collection, and statistical inference (Garfield, 2003). The Statistics Concepts Inventory (SCI) was developed to assess statistical understanding but it was written for a specific audience of engineering students in statistics (Reed-Rhoads, Murphy, & Terry, 2006). Garfield, delMas, and Chance (2002) aimed to develop an assessment instrument that would have broader coverage of both the statistical content typically covered in the first, non-mathematical statistics course, and would apply to the broader range of students who enroll in these courses.

2. THE ARTIST PROJECT

The National Science Foundation (NSF) funded the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project (DUE-0206571) to address the assessment challenge in statistics education as presented by Garfield and Gal (1999), who outlined the need to develop reliable, valid, practical, and accessible assessment items and instruments. The ARTIST Web site (https://app.gen.umn.edu/artist/) now provides a wide variety of assessment resources for evaluating students' statistical literacy (e.g., understanding words and symbols, being able to read and interpret graphs and terms), statistical reasoning (e.g., reasoning with statistical information), and statistical thinking (e.g., asking questions and making decisions involving statistical information). These resources were designed to assist faculty who teach statistics across various disciplines (e.g., mathematics, statistics, and psychology) in assessing student learning of statistics, to better evaluate individual student achievement, to evaluate and improve their courses, and to assess the impact of reform-based instructional methods on important learning outcomes.

3. DEVELOPMENT OF THE CAOS TEST

An important component of the ARTIST project was the development of an overall Comprehensive Assessment of Outcomes in Statistics (CAOS). The intent was to develop a reliable assessment consisting of a set of items that students completing any introductory statistics course would be expected to understand. Given that a reliable assessment could be developed, a second goal was to identify areas where students do and do not make significant gains in their statistical understanding and reasoning.

The CAOS test was developed through a three-year iterative process of acquiring existing items from instructors, writing items for areas not covered by the acquired items, revising items, obtaining feedback from advisors and class testers, and conducting two large content validity assessments. During this process the ARTIST team developed and revised items and the ARTIST advisory board provided valuable feedback as well as validity ratings of items, which were used to determine and improve content validity for the targeted population of students (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

The ARTIST advisory group initially provided feedback and advice on the nature and content of such a test. Discussion led to the decision to focus the instrument on different aspects of reasoning about variability, which was viewed as the primary goal of a first course. This included reasoning about variability in distributions, in comparing groups, in sampling, and in sampling distributions. The ARTIST team had developed an online assessment item database with over 1000 items as part of the project. Multiple choice items to be used in the CAOS test were initially selected from the ARTIST item database or were created. All items were revised to ensure they involved real or realistic contexts and data, and to ensure that they followed established guidelines for writing multiple choice items (Haladyna, Downing, & Rodriguez, 2002). The first set of items was evaluated by the ARTIST advisory group, who provided ratings of content validity and identified important concepts that were not measured by the test. The ARTIST team revised the test and created new items to address missing content. An online prototype of CAOS was developed during summer 2004, and the advisors engaged in another round of validation and feedback in early August, 2004. This feedback was then used to produce the first version of CAOS, which consisted of 34 multiple-choice items. This version was used in a pilot study with introductory statistics students during fall 2004. Data from the pilot study were used to make additional revisions to CAOS, resulting in a second version of CAOS that consisted of 37 multiple choice items.

The second version, called CAOS 2, was ready to launch as an online test in January 2005. Administration of the online test required a careful registration of instructors, a means for students to securely access the test online, and provision for instructors to receive timely feedback of test results. In order to access the online tests, an instructor requested an access code, which was then used by students to take the test online. As soon as the students completed the test, either in class or out of class, the instructor could download two reports of students' data. One was a copy of the test, with percentages filled in for each response given by students, and with the correct answers highlighted. The other report was a spreadsheet with the total percentage correct score for each student.

3.1. CLASS TESTING OF CAOS 2

The first large scale class testing of the online instruments was conducted during spring 2005. Invitations were sent to teachers of high school Advanced Placement (AP) and college statistics courses through e-mail lists (e.g., AP community, Statistical Education Section of the American Statistics Association). In order to gather as much data as possible, a hard copy version of the test with machine readable bubble sheets was also offered. Instructors signed up at the ARTIST Web site to have their students take CAOS 2 as a pretest and /or a posttest, using either the online or bubble sheet format.

Many instructors registered their students to take the ARTIST CAOS 2 test as a pretest at the start of a course and as a posttest toward the end of the course. Although it was originally hoped that all tests would be administered in a controlled classroom setting, many instructors indicated the need for out-of-class testing. Information gathered from registration forms also indicated that instructors used the CAOS results for a variety of purposes, namely, to assign a grade in the course, for review before a course exam, or to assign extra credit. Nearly 100 secondary-level students and 800 college-level students participated. Results from the analysis of the spring 2005 data were used to make additional changes, which produced a third version of CAOS (CAOS 3).

3.2. EVALUATION OF CAOS 3 AND DEVELOPMENT OF CAOS 4

The third version of CAOS (CAOS 3) was given to a group of 30 statistics instructors who were faculty graders of the Advanced Placement Statistics exam in June 2005, for another round of validity ratings. Although the ratings indicated that the test was measuring what it was designed to measure, the instructors also made many suggestions for changes. This feedback was used to add and delete items from the test, as well as to make extensive revisions to produce a final version of the test, called CAOS 4, consisting of 40 multiple choice items. CAOS 4 was administered in a second large scale testing during fall 2005. Results from this large scale, national sample of college-level students are reported in the following sections.

In March 2006, a final analysis of the content validity of CAOS 4 was conducted. A group of 18 members of the advisory and editorial boards of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) were used as expert raters. These individuals are statisticians who are involved in teaching statistics at the college level, and who are considered experts and leaders in the national statistics education community. They were given copies of the CAOS 4 test that had been annotated to show what each item was designed to measure. After reviewing the annotated test, they were asked to respond to a set of questions about the validity of the items and instrument for use as an outcome measure of student learning after a first course in statistics. There was unanimous agreement by the expert raters with the statement “CAOS measures basic outcomes in statistical literacy and reasoning that are appropriate for a first course in statistics,” and 94% agreement with the statement “CAOS measures important outcomes that are common to most first courses in statistics.” In addition, all raters agreed with the statement “CAOS measures outcomes for which I would be disappointed if they were not achieved by students who succeed in my statistics courses.” Although some raters indicated topics that they felt were missing from the scale, there was no additional topic identified by a majority of the raters. Based on this evidence, the assumption was made that CAOS 4 is a valid measure of important learning outcomes in a first course in statistics.

4. CLASS TESTING OF CAOS 4

4.1. DESCRIPTION OF THE SAMPLE

In the fall of 2005 and spring of 2006, CAOS 4 was administered as an online and hard copy test for a final round of class testing and data gathering for psychometric analyses. The purpose of the study was to gather baseline data for psychometric analysis and not to conduct a comparative study (e.g., performance differences between traditional and reform-based curricula). The recruitment approach used for class testing of CAOS 2 was employed, as well as inviting instructors who had given previous versions of CAOS to participate. A total of 1944 students completed CAOS 4 as a posttest. Several criteria were used to select students from this larger pool as a sample with which to conduct a reliability analysis of internal consistency. To be included in the sample, students had to respond to all 40 items on the test and either have completed CAOS 4 in an in-class, controlled setting or, if the test was taken out of class, have taken at least 10 minutes, but no more than 60 minutes, to complete the test. The latter criterion was used to eliminate students who did not engage sufficiently with the test questions or who spent an excessive amount of time on the test, possibly looking up answers. In addition, students enrolled in high school AP courses were not included in the analysis. Analysis of data from earlier versions of the CAOS test produced significant differences in percentage correct when the AP and college samples were compared. Inclusion of data from AP students might produce results that are not representative of the general undergraduate population, and a comparison of high school AP and college students is beyond the scope of this study.

A total of 1470 introductory statistics students, taught by 35 instructors from 33 higher education institutions from 21 states across the United States met these criteria and were included in the sample (see Table 1). The majority of the students whose data were used for the reliability analysis were enrolled at a university or a four-year college, with about one fourth of the students enrolled in two-year or technical colleges. A little more than half of the students (57%) were females, and 74% of the students were Caucasian.

Table 1. Number of higher education institutions, instructors, and students per institution type for students who completed the CAOS 4 posttest

Institution Type	Number of institutions	Number of instructors	Number of students	Percent of students
2-year/technical	6	6	341	23.1
4-year college	13	14	548	37.3
University	14	15	581	39.5
Total	33	35	1470	

Table 2 shows the mathematics requirements for entry into the statistics course in which students enrolled. The largest group was represented by students in courses with a high school algebra requirement, followed by a college algebra requirement and no mathematics requirement, respectively. Only 3% of the students were enrolled in a course with a calculus prerequisite.

The majority of the students (64%) took the CAOS 4 posttest in class (henceforth refered to as CAOS). Only four instructors used the CAOS test results as an exam score, which accounted for 12% of the students. The most common uses of the CAOS posttest results were to assign extra credit (35%), or for review prior to the final exam (19%), or both (13%).

Table 2. Number and percent of students per course type

Mathematics prerequisite	Number of students	Percent of students
No mathematics requirement	398	27.1
High school algebra	611	41.6
College algebra	420	28.6
Calculus	41	2.8

4.2. RELIABILITY ANALYSIS

Using the sample of students described above, an analysis of internal consistency of the 40 items on the CAOS posttest produced a Cronbach's alpha coefficient of 0.82. Different standards for an acceptable level of reliability have been suggested, with lower limits ranging from 0.5 to 0.7 (see Pedhazur & Schmelkin, 1991). The CAOS test was judged to have acceptable internal consistency for students enrolled in college-level, non-mathematical introductory statistics courses given that the estimated internal consistency reliability is well above the range of suggested lower limits.

5. ANALYIS OF PRETEST TO POSTTEST CHANGES

A major question that needs to be addressed is whether students enrolled in a first statistics course make significant gains from pretest to posttest on the CAOS test. The total percentage correct scores from a subset of students who completed CAOS as both a pretest (at the beginning of the course) and as a posttest (at the end of the course) were compared for 763 introductory statistics students.

5.1. DESCRIPTION OF THE SAMPLE

The 763 students in this sample of matched pretests and posttests were taught by 22 instructors at 20 higher education institutions from 14 states across the United States (see Table 3). Students from four-year colleges made up the largest group, followed closely by university students. Eighteen percent of the students were from two-year or technical colleges. The majority of the students were females (60%), and 77% of the students were Caucasian.

Table 4 shows the distribution of mathematics requirements for entry into the statistics courses in which students enrolled. The largest group was represented by students in courses with a high school algebra requirement, followed by no mathematics requirement, and a college algebra requirement, respectively. Only about 4% of the students were enrolled in a course with a calculus prerequisite.

Table 3. Number of higher education institutions, instructors, and students per institution type for students who completed both a pretest and a posttest

Institution Type	Number of institutions	Number of instructors	Number of students	Percent of students
2-year/technical	4	4	138	18.1
4-year college	10	11	395	51.8
University	6	7	230	30.1
Total	20	22	763	

Table 4. Number and percent of students per type of mathematics prerequisite

Mathematics Prerequisite	Number of students	Percent of students
No mathematics requirement	197	25.8
High school algebra	391	51.2
College algebra	161	21.1
Calculus	14	1.8

Sixty-six percent of the students received the CAOS posttest as an in-class administration, with the remainder taking the test online outside of regularly scheduled class time. Only four instructors used the CAOS posttest scores solely as an exam grade in the course, which accounted for 11% of the students. The most common use of the CAOS posttest results for students who took both the pretest and posttest was to assign extra credit (23% of the students). For 22% of the students the CAOS posttest was used only for review, whereas another 16% received extra credit in addition to using CAOS as a review before the final exam. For the remainder of the students (29%), instructors indicated some other use such as program or course evaluation.

5.2. PRETEST TO POSTTEST CHANGES IN CAOS TEST SCORES

There was an increase from an average percentage correct of 44.9% on the pretest to an average percentage correct of 54.0% on the posttest (se = 0.433; $t(762) = 20.98$, $p < 0.001$). Although statistically significant, this was only a small average increase of 9 percentage points (95% CI = [8.2,9.9] or 3.3 to 4.0 of the 40 items). It was surprising to find that students were correct on little more than half of the items, on average, by the end of the course. To further investigate what could account for the small gain, student responses on each item were compared to see if there were items with significant gains, items that showed no improvement, or items where the percentage of students with correct answers decreased from pretest to posttest.

6. PRETEST TO POSTTEST CHANGES FOR INDIVIDUAL ITEMS

The next step in analyzing pretest to posttest gains was to look at changes in correct responses for individual items. Matched-pairs *t* tests were conducted for each CAOS item to test for statistically significant differences between pretest and posttest percentage correct. Responses to each item on the pretest and posttest were coded as 0 for an incorrect response and 1 for a correct response. This produced four different response patterns across the pretest and posttest for each item. An “incorrect” response pattern consisted of an incorrect response on both the pretest and the posttest. A “decrease” response pattern was one where a student selected a correct response on the pretest and an incorrect response on the posttest. An “increase” response pattern occurred when a student selected an incorrect response on the pretest and a correct response on the posttest. A “pre & post” response pattern consisted of a correct response on both the pretest and the posttest. The percentage of students who fell into each of these response pattern categories is given in Appendix A.

The change from pretest to posttest in the percentage of students who selected the correct response was determined by the difference between the percentage of students who fell into the “increase” and “decrease” categories. This is a little more apparent if it is recognized that the percentage of students who gave a correct response on the pretest was equal to the percentage in the “decrease” category plus the percentage in the “pre & post” category. Similarly, the percentage of students who gave a correct response on the posttest was equal to the percentage in the “increase” category added to the percentage in the “pre & post” category. When the percentage of students in the “decrease” and “increase” categories were about the same, the change tended to not produce a statistically significant effect relative to sampling error. When there was a large difference in the percentage of students in these two categories (e.g., one category had twice or more students than the other category), the change had the potential to produce a statistically significant effect relative to sampling error. Comparison of the percentage of students in these two “change” categories can be used to interpret the change in percentage from pretest to posttest.

A per test Type I Error limit was set at $\alpha_c = 0.001$ to keep the study-wide Type I Error rate at $\alpha = 0.05$ or less across the 46 paired *t* tests conducted (see Tables 5 through 9). For each CAOS item that produced a statistically significant change from pretest to posttest, multivariate analyses of variance (MANOVA) were conducted. The dependent variables for each analysis consisted of a 0/1 coded response for a particular item on the pretest and the posttest (0 = incorrect, 1 = correct). The two independent variables for each MANOVA consisted of the pretest/posttest repeated measure and either type of institution or type of mathematics prerequisite. Separate MANOVAs were conducted using only one of the two between-subjects grouping variables because the two variables were not completely crossed. A p-value limit of 0.001 was again used to control the experiment-wise Type I Error rate. If no interaction was found with either variable, an additional MANOVA was conducted using instructor as a grouping variable, to see if a statistically significant change from pretest to posttest was due primarily to large changes in only a few classrooms.

The following sections describe analyses of items that were grouped into the following categories: (a) those that had high percentages of students with correct answers on both the pretest and the posttest, (b) those that had moderate percentages of correct answers on both pretest and posttest, (c) those that showed the largest increases from pretest to posttest, and (d) those that had low percentages of students with correct responses on both the pretest and the posttest. Tables 5 through 8 present a brief description of what each item assessed, report the percentage of students who selected a correct response separately for the pretest and the posttest, and indicate the p-value of the respective matched-pairs *t* statistic for each item.

6.1. ITEMS WITH HIGH PERCENTAGES OF STUDENTS WITH CORRECT RESPONSES ON BOTH PRETEST AND POSTTEST

It was surprising to find several items on which students provided correct answers on the pretest as well as on the posttest. These were eight items on which 60% or more of the students demonstrated an ability or conceptual understanding at the start of the course, and on which 60% or more of the students made correct choices at the end of the course (Table 5). A majority of the students were correct on both the pretest and the posttest for this set of items. Across the eight items represented in Table 5, about the same percentage of students (between 5% and 21%) had a decrease response pattern as had an increase response pattern for each item, with the exceptions of items 13 and 21 (see Appendix A). The net result was that the change in percentage of students who were correct did not meet the criterion for statistical significance for any of these items.

Table 5. Items with 60% or more of students correct on the pretest and the posttest

			% of Students Correct		Paired *t*
Item	Measured Learning Outcome	*n*	Pretest	Posttest	*p*
1	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.	760	71.5	73.6	0.266
11	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities.	756	88.0	88.2	0.856
12	Ability to compare groups by comparing differences in averages.	753	85.3	85.8	0.741
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.	752	61.8	73.5	<0.001
18	Understanding of the meaning of variability in the context of repeated measurements, and in a context where small variability is desired.	746	80.6	80.6	1.00
20	Ability to match a scatterplot to a verbal description of a bivariate relationship.	748	90.5	92.5	0.132
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).	749	73.6	83.7	<0.001
23	Understanding that no statistical significance does not guarantee that there is no effect.	735	63.1	64.4	0.588

Around 70% of the students were able to select a correct description and interpretation of a histogram that included a reference to the context of the data (item 1). The most common mistake on the posttest was to select the option that correctly described shape, center, and spread, but did not provide an interpretation of these statistics within the context of the problem.

In general, students demonstrated facility on both the pretest and posttest with using distributional reasoning to make comparisons between two groups (items 11, 12, and 13). Almost 90% of the students on the pretest and posttest correctly indicated that comparisons based on single cases were not valid. Students had a little more difficulty with item 13, which required the knowledge that comparing groups does not require equal sample sizes in each group, especially if both sets of data are large. Students appear to have good informal intuitions or understanding of how to compare groups. However, the belief that groups must be of equal size to make valid comparisons is a persistent misunderstanding for some students.

A majority of students on the pretest appeared to understand that statistical significance does not mean that there is no effect (item 23). However, making a correct choice on this item was not as persistent as for the items described above; a little more than a third of the students did not demonstrate this understanding on the posttest.

6.2. ITEMS THAT SHOWED INCREASES IN PERCENTAGE OF STUDENT WITH CORRECT RESPONSES FROM PRETEST TO POSTTEST

There were seven items on which there was a statistically significant increase from pretest to posttest, and at least 60% of the students made a correct choice on the posttest (Table 6). For all seven items, less than half of the students were correct on both the pretest and the posttest (see Appendix A). Whereas between 6% and 16% of the students had a decrease response pattern across the items, there were two to five times as many students with an increase response pattern for each item, with the exception of item 34. This resulted in statistically significant increases from pretest to posttest in the percentage of students who chose correct responses for each item.

Around half of the students on the pretest were able to match a histogram to a description of a variable expected to have a distribution with a negative skew (item 3), a variable expected to have a symmetric, bell-shaped distribution (item 4), and a variable expected to have a uniform distribution (item 5), with increases of about 15 percentage points from pretest to posttest for each of the three items. About half of the students correctly indicated that a small p-value is needed to establish statistical significance (item 19), and this increased by 23 percentage points on the posttest. A significant interaction was produced for pretest to posttest change by instructor ($F(21, 708) = 2.946$, $p < 0.001$). Three instructors had a decrease of seven to 23 percentage points from pretest to posttest, one instructor had essentially no change, 11 instructors had an increase of 10 to 28 percentage points, and seven instructors had an increase of 36 to 63 percentage points. Five of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and others with relatively large increases. Overall, the majority of instructors (19 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

On the pretest, only one third of the students recognized an invalid interpretation of a confidence interval as the percentage of the population data values between the confidence limits (item 29), which increased to around two thirds on the posttest. There was a statistically significant interaction with instructor [$F(21,703) = 3.163$, $p<.001$]. There was essentially no change in percentage correct from pretest to posttest for three of the instructors. For the other 19 instructors, students showed an increase of 23 to 60 percentage points. The instructor with the highest increase was not the same instructor with the highest increase for item 19. The increase was statistically significant at $p<.001$ for the students of only nine instructors, which could account for the interaction.

Table 6. Items with 60% or more of students correct on the posttest and statistically significant gain

			% of Students Correct		Paired *t*
Item	Measured Learning Outcome	*n*	Pretest	Posttest	*p*
3	Ability to visualize and match a histogram to a description of a variable (negatively skewed distribution for scores on an easy quiz).	760	56.7	73.2	<0.001
4	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).	757	48.0	63.1	<0.001
5	Ability to visualize and match a histogram to a description of a variable (uniform distribution for the last digit of phone numbers sampled from a phone book).	758	55.9	71.1	<0.001
19	Understanding that low *p*-values are desirable in research studies.	730	49.9	68.5	<0.001
29	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).	725	32.6	67.6	<0.001
31	Ability to correctly interpret a confidence interval.	720	47.1	74.3	<0.001
34	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.	724	55.3	65.2	<0.001

About half of the students recognized a valid interpretation of a confidence interval on the pretest (item 31), which increased to three fourths on the posttest. There was a statistically significant interaction with instructor [$F(21,698) = 2.787$, $p<.001$]. Students of 20 of the instructors had an increase of 23 to 60 percentage points from pretest to posttest. The students of the other two instructors had a decrease of 7 and 15 percentage changes, respectively, neither of which were statistically significant. The instructor with the highest increase was not the same instructor with the highest increase for either item 19 or item 29. The increase was statistically significant at $p<.001$ for the students of only six instructors, which could account for the interaction.

Finally, athough a little more than half of the students could correctly identify a plausible random sample taken from a population on the pretest, this increased by 10 percentage points on the posttest (item 34). Whereas these students showed both practical and statistically significant gains on all of the items in Table 6, anywhere from 26% to 37% still did not make the correct choice for this set of items on the posttest.

There were thirteen additional items that produced statistically significant increases in percentage correct from pretest to posttest, but where the percentage of students with correct responses on the posttest was still below 60% (Table 7). Similar to the items in Table 6, between 7% and 18% of the students had a decrease response pattern. However, for each item, about one and a half to three times as many students had a response pattern that qualified as an increase. The net result was a statistically significant increase in the percentage of students correct for all thirteen items.

In general, students demonstrated some difficulty interpreting graphic representations of data. Item 2 asked students to identify a boxplot that represented the same data displayed in a histogram. Performance was around 45% of students correct on the pretest with posttest performance just under 60%. On item 6, less than one fourth of the students on the pretest and the posttest demonstrated the understanding that a graph like a histogram is needed to show shape, center, and spread of a distribution of quantitative data. The 10 percentage point increase from pretest to posttest in percentage of students selecting the correct response was statistically significant. Most students (43% on the pretest and 53% on the posttest) selected a bar graph with a bell shape, but such a graph cannot be used to directly determine the mean, variability, and shape of the measured variable. Students demonstrated a tendency to select an apparent bell-shaped or normal distribution, even when this did not make sense within the context of the problem.

The MANOVAs conducted for item 6 responses with type of institution and type of mathematics preparation did not produce significant interactions. The MANOVA that included instructor as an independent variable did produce a statistically significant interaction between pretest to posttest change and instructor ($F(21, 732) = 3.224$, $p < 0.001$). Only one of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. Two instructors had a small decrease in percentage of students correct from pretest to posttest, three instructors had essentially no change, 12 instructors had an increase of seven to 18 percentage points, and five instructors had an increase of 26 to 47 percentage points. The differential increase in percentage of students who gave a correct response may account for the interaction. Overall, the general trend was for an increase in the percentage of students with correct responses to item 6.

A very small percentage of students demonstrated a correct understanding of the median in the context of a boxplot (item 10) on the pretest, with about a 9% improvement on the posttest. Item 10 presented two boxplots positioned one above the other on the same scale. Both boxplots had the same median and roughly the same range. The width of the box for one graph was almost twice the width of the other graph, with consequently shorter whiskers. On the posttest, most students (66%) chose a response that indicated that the boxplot with a longer upper whisker would have a higher percentage of data above the median. A significant interaction was produced for pretest to posttest change by instructor ($F(21, 732) = 3.958$, $p < 0.001$). Only one of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. Five instructors had a decrease of six to 14 percentage points from pretest to posttest, two instructors had essentially no change, nine instructors had an increase of five to 17 percentage points, and six instructors had an

Table 7. Items with less than 60% of students correct on the posttest, gain statistically significant

			% of Students Correct		Paired *t*
Item	Measured Learning Outcome	*n*	Pretest	Posttest	*p*
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram).	759	45.5	56.3	<0.001
6	Understanding that to properly describe the distribution (shape, center, and spread) of a quantitative variable, a graph like a histogram is needed.	754	15.1	25.2	<0.001
10	Understanding of the interpretation of a median in the context of boxplots.	754	19.6	28.3	<0.001
14	Ability to correctly estimate and compare standard deviations for different histograms. Understands lowest standard deviation would be for a graph with the least spread (typically) away from the center.	746	34.3	51.7	<0.001
15	Ability to correctly estimate standard deviations for different histograms. Understands highest standard deviation would be for a graph with the most spread (typically) away from the center.	747	38.3	46.9	<0.001
16	Understanding that statistics from small samples vary more than statistics from large samples.	747	22.8	31.9	<0.001
17	Understanding of expected patterns in sampling variability.	746	42.8	50.3	<0.001
27	Ability to recognize an incorrect interpretation of a p-value (prob. treatment is effective).	717	42.3	52.7	<0.001
30	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits).	723	31.4	44.2	<0.001
35	Ability to select an appropriate sampling distribution for a population and sample size.	719	34.5	44.2	<0.001
38	Understanding of the factors that allow a sample of data to be generalized to the population.	715	26.0	37.9	<0.001
39	Understanding of when it is not wise to extrapolate using a regression model.	710	17.9	24.5	0.001
40	Understanding of the logic of a significance test when the null hypothesis is rejected.	716	41.9	52.0	<0.001

increase of 31 to 61 percentage points. Again, the differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and others with relatively large increases. Overall, the majority of instructors (15 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

Item 14 asked students to determine which of several histograms had the lower standard deviation. A little over half of the students answered this item correctly on the posttest. The 17 percentage point increase in percentage correct from pretest to posttest, however, was statistically significant.

Item 15 asked students to determine which of several histograms had the highest standard deviation. Similar to item 14, a little under half of the students answered this item correctly on the posttest. There was about a nine percent increase in percentage correct from pretest to posttest. A significant interaction was found for pretest to posttest change by course type ($F(3, 743) = 5.563$, $p < 0.001$). Simple effects analyses indicated that the change from pretest to posttest was statistically significant increase for students in courses with no mathematics prerequisite ($F(1, 189) = 10.851$, $p = 0.001$) or a high school algebra prerequisite ($F(1, 383) = 16.460$, $p < 0.001$), but not for students in courses with college algebra ($F(1, 158) = 1.872$ $p = 0.173$) or calculus ($F(1, 13) = 1.918$, $p = 0.189$) prerequisites. In fact, the percentage of students correct on item 15 decreased for the latter two groups, although the differences were not statistically significant.

Item 16 required the understanding that statistics from relatively small samples vary more than statistics from larger samples. Although the increase was statistically significant ($p < 0.001$), only about one fifth of the students answered this item correctly on the pretest and less than a third did so on the posttest. A slight majority of students on the posttest indicated that both sample sizes had the same likelihood of producing an extreme value for the statistic. A significant interaction was found for pretest to posttest change by type of institution ($F(2, 744) = 7.169$, $p < 0.001$). Simple effects analyses (Howell, 2002) did not produce a significant effect for type of institution on the pretest ($F(2, 1292) = 2.701$, $p = 0.068$), but the effect was significant on the posttest ($F(2, 1292) = 9.639$, $p < 0.001$). Thirty-six percent of students enrolled at a four-year college and 34% of those attending a university gave a correct response on the posttest, whereas only 17% of those enrolled in a technical or two-year college gave a correct response. The percentage of students who gave a correct response was about the same on the pretest and posttest for technical and two-year college students, whereas four-year colleges had a gain of 9 percentage points and universities had a gain of 16 percentage points. Overall, the change in percentage of students who were correct on item 16 was primarily due to students enrolled at four-year institutions and universities.

Item 17 presented possible results for five samples of equal sample size taken from the same population. Less than half the students on the pretest and posttest chose the sequence that represented the expected sampling variability in the sample statistic. About one third of students on the pretest (36%) and the posttest (33%) indicated that all three sequences of sample statistics were just as plausible, even though one sequence showed an extreme amount of sampling variability given the sample size, and another sequence presented the same sample statistic for each sample (i.e., no sampling variability). In addition, 74% of the students who gave an erroneous response to item 17 on the posttest also selected an erroneous response for item 16.

There was a statistically significant ($p < 0.001$) increase from pretest to posttest in the percentage of students who indicated that the confidence level indicated the percentage of all sample means that fall between the confidence limits (item 30). However, the percentage went from 31% on the pretest to 44% on the posttest, so that the majority of students did not indicate this understanding by the end of their statistics courses. A significant interaction was produced for pretest to posttest change by instructor (*F*(21, 701) = 2.237, *p* < 0.001). Two instructors had a decrease of 10 and 43 percentage points, respectively, from pretest to posttest, one instructor had essentially no change, 17 instructors had an increase between four and 16 percentage points, and two instructors had an increase of 21 and 37 percentage points, respectively. Three of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at *p* < 0.001. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and others with relatively large increases. Overall, the majority of instructors (19 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

Item 27 presented a common misinterpretation of a p-value as the probability that a treatment is effective. Forty percent of the students answered correctly on the pretest that the statement was invalid, which increased to 53% on posttest. Although the increase was statistically significant, nearly half of the students indicated that the statement was valid at the end of their respective courses.

Item 35 asked students to select a graph from among three histograms that represented a sampling distribution of sample means for a given sample size. Slightly more than one third did so correctly on the pretest, with 10% more students selecting the correct response on the posttest.

Many students did not demonstrate a good understanding of sampling principles. Only one fifth of the students on the pretest, and nearly 40% on the posttest made a correct choice of conditions that allow generalization from a sample to a population (item 38). Even though this was a statistically significant gain from pretest to posttest, over 62% indicated that a random sample of 500 students presented a problem for generalization on the posttest (supposedly because it was too small a sample to represent the 5000 students living on campus). No statistically significant interactions were produced by the MANOVA analyses.

Only one fifth of the students indicated on the posttest that it is not appropriate to extrapolate a regression model to values of the predictor variable that are well beyond the range of values investigated in a study (item 39). A significant interaction was produced for pretest to posttest change by instructor (*F*(21, 688) = 4.881, *p* < 0.001). Two of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced statistically significant differences at *p* < 0.001. The two instructors were both from four-year institutions and had increases of 40 and 61 percentage points, respectively. Among the other instructors, four had a decrease of five to 16 percentage points from pretest to posttest, five instructors had essentially no change (between a decrease of five to an increase of five percentage points), seven instructors had an increase of six to 19 percentage points, and four instructors had an increase of 23 to 30 percentage points. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and a few with relatively large increases. Overall, the majority of instructors (13 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

About half of the students could identify a correct interpretation of rejecting the null hypothesis (item 40) on the posttest. Although there was a statistically significant gain in correct responses from pretest to posttest, about one third of the students indicated that rejecting the null hypothesis meant that it was definitely false, which was five percentage points higher than the percentage who gave this response on the pretest. A significant interaction was produced for pretest to posttest change by instructor (*F*(21, 694) = 2.392, $p < 0.001$). Two of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced statistically significant differences at $p < 0.001$. The two instructors were both from four-year institutions and had increases of 39 and 55 percentage points, respectively. Among the other instructors, five had a decrease of five to 22 percentage points from pretest to posttest, five instructors had essentially no change (between a decrease of five to an increase of 4 percentage points), nine instructors had an increase of six to 14 percentage points, and three instructors had an increase of 21 to 39 percentage points. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and a few with relatively large increases. Overall, the majority of instructors (13 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

6.3. ITEMS WITH LOW PERCENTAGES OF STUDENTS WITH CORRECT RESPONSES ON BOTH THE PRETEST AND THE POSTTEST

Table 8 shows that for a little less than one third of the items on the CAOS test less than 60% of the students were correct on the posttest with the change from pretest to posttest not statistically significant, despite having experienced the curriculum of a college-level first course in statistics. Across all of these items, similar percentages of students (between 6% and 30%) had a decrease response pattern as had an “increase” response pattern (see Appendix A). The overall result was that none of the changes from pretest to posttest in percentage of students selecting a correct response were statistically significant.

Students had very low performance, both pretest and posttest, on item 7, which required an understanding for the purpose of randomization (to produce treatment groups with similar characteristics). On the posttest, about 30% of the students chose “to increase the accuracy of the research results,” and another 30% chose “to reduce the amount of sampling error.”

Students demonstrated some difficulty with understanding how to correctly interpret boxplots. Items 8 and 9 were based on the same two boxplots presented for item 10 (Table 7). Item 8 asked students to identify which boxplot represented a distribution with a larger standard deviation. One boxplot had a slightly larger range (difference of approximately five units) with an interquartile range that was about twice as large as the interquartile range for the other boxplot. Around 59% of the students chose this graph to have a larger standard deviation on the posttest. On item 9, only one fifth of the students demonstrated an understanding that boxplots do not provide estimates for percentages of data above or below values except for the quartiles. The item asked students to indicate which of the two boxplots had a greater percentage of cases at or below a specified value. The value did not match any of the quartiles or extremes marked in either boxplot, so the correct response was that it was impossible to determine. Given that item 9 has four response choices, the correct response rate was close to chance level on both the pretest and posttest. Fifty-eight percent of students on the posttest indicated that the boxplot with the longer lower whisker had a higher percentage of cases below the indicated value, similar to the erroneous response to item 10. On the posttest, 48% of the students selected the identified erroneous responses to both items 9 and 10.

Table 8. Items with less than 60% of students correct on the posttest, gain not statistically significant

			% of Students Correct		Paired *t*
Item	Measured Learning Outcome	*n*	Pretest	Posttest	*p*
7	Understanding of the purpose of randomization in an experiment.	754	8.5	12.3	0.010
8	Ability to determine which of two boxplots represents a larger standard deviation.	755	54.7	59.2	0.060
9	Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles.	751	23.3	26.6	0.100
22	Understanding that correlation does not imply causation.	743	54.6	52.6	0.371
24	Understanding that an experimental design with random assignment supports causal inference.	731	58.5	59.5	0.689
25	Ability to recognize a correct interpretation of a p-value.	712	46.8	54.5	0.004
26	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is not effective).	719	53.1	58.6	0.038
28	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits).	729	48.4	43.2	0.029
32	Understanding of how sampling error is used to make an informal inference about a sample mean.	718	16.9	17.1	0.883
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left.	730	41.5	39.7	0.477
36	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.	719	52.7	53.0	0.909
37	Understanding of how to simulate data to find the probability of an observed value.	722	20.4	19.5	0.659

Although it was noted earlier that students could correctly identify a scatterplot given a description of a relationship between two variables, they did not perform as well on another item related to interpreting correlation. About one third (36%) of the students chose a response indicating that a statistically significant correlation establishes a causal relationship (item 22). Item 24 required students to understand that causation can be inferred from a study with an experimental design that uses random assignment to treatments. The percentage of students answering this item correctly on the posttest was just below the threshold of 60%.

Items 25 and 26 measured students' ability to recognize a correct and an incorrect interpretation of a p-value, respectively. There was a noticeable change from pretest to posttest in the percentage of students indicating that item 25 was a valid interpretation, but the difference was just above the threshold for statistical significance. About 55% of the students answered item 5 correctly and 59% answered item 26 correctly on the posttest. Results for these two items, along with item 27, indicate that many students who identified a correct interpretation of a p-value as valid also indicated that an incorrect interpretation was valid. In fact, of the 387 students who answered item 25 correctly on the posttest, only 5% also indicated that the statements for items 26 and 27 were invalid. For the remainder of these students, 56% thought one of the incorrect interpretations was valid, and 39% indicated both incorrect interpretations as valid.

Students did not demonstrate a firm grasp of how to interpret confidence intervals. There was an increase in the percentage of students who incorrectly indicated that the confidence level represents the expected percentage of sample values between the confidence limits (item 28), although the difference was not statistically significant.

An item related to sampling variability proved difficult for students. Item 32 required students to recognize that an estimate of sampling error was needed to conduct an informal inference about a sample mean. Less than 20% of the students made a correct choice on the pretest and posttest. A slight majority of the students (54% pretest, 59% posttest) chose the option that based the inference solely on the sample standard deviation, not taking sample size and sampling variability into account.

Item 33 required the understanding that a distribution with a median greater than the mean is most likely skewed to the left. There was a decrease, though not statistically significant, in the number of students who demonstrated this understanding. The percentage of those who incorrectly selected a somewhat symmetric, mound-shaped bar graph increased from 54% on the pretest to 59% on the posttest. Sixty-four percent of those who made this choice on the posttest also incorrectly chose the bell-shaped bar graph for item 6 (Table 7) discussed earlier.

A little more than half of the students correctly indicated that ratios based on marginal totals were needed to make comparisons between rows in a two-way table of counts (item 36). One third of the students incorrectly chose proportions based on the overall total count on the posttest.

Eighty percent of the students did not demonstrate knowledge of how to simulate data to estimate the probability of obtaining a value as or more extreme than an observed value (item 37). In a situation where a person has to predict between two possible outcomes, the item asked for a way to determine the probability of making at least four out of six correct predictions just by chance. On the posttest, 46% of the students indicated that repeating the experiment a large number of times with a single individual, or repeating the experiment with a large group of people and determining the percentage who make four out of six correct predictions, were equally effective as calculating the percentage of sequences of six trials with four or more correct predictions for a computer simulation with a 50% chance of a correct prediction on each trial.

6.4. ITEM RESPONSES THAT INDICATED INCREASED MISCONCEPTIONS AND MISUNDERSTANDINGS

Whereas some of the items discussed in the previous section showed a drop in the percentage of students with correct responses from pretest to posttest, none of these differences was statistically significant. There were, however, several items with noticeable increases from pretest to posttest in the percentage of students selecting a specific erroneous response (Table 9). The change in percentage of students with correct responses was statistically significant for four of the six items in Table 9. None of these responses produced statistically significant interactions between pretest to posttest increases and either type of institution, type of mathematics preparation, or instructor. Most of these misunderstandings and misconceptions were discussed in earlier presentations of the results. They include selecting a bell-shaped bar graph to represent the distribution of a quantitative variable (item 6), confusing random assignment with random sampling (item 7), selecting a histogram with a larger number of different values as having a larger standard deviation (item 15), inferring causation from correlation (item 22), use of grand totals to calculate conditional probabilities (item 36), and indicating that rejecting the null hypothesis means the null hypothesis is definitely false (item 40).

Table 9. Items with an increase in a misconception or misunderstanding from pretest to posttest

			% of Students		Paired *t*
Item	Misconception or Misunderstanding	*n*	Pretest	Posttest	*p*
6	A bell-shaped bar graph to represent the distribution for a quantitative variable.	754	43.0	52.8	<0.001
7	Random assignment is confused with random sampling or thinks that random assignment reduces sampling error.	754	36.2	49.2	<0.001
15	When comparing histograms, the graph with the largest number of different values has the larger standard deviation (spread not considered).	747	26.5	33.1	0.002
22	Causation can be inferred from correlation.	743	27.1	35.9	<0.001
36	Grand totals are used to calculate conditional probabilities.	719	25.2	33.4	<0.001
40	Rejecting the null hypothesis means that the null hypothesis is definitely false.	716	26.7	32.4	0.015

Across this set of items, 13% to 17% of the students had a decrease response pattern with respect to the identified erroneous response (see Appendix B). For each item, between one and a half to two times as many students had an increase response pattern with respect to giving the erroneous response. The result was a statistically significant increase in the percentage of students selecting the identified responses for four of the items. Together, these increases indicate that a noticeable number of students developed misunderstandings or misconceptions by the end of the course that they did not demonstrate at the beginning.

7. DISCUSSION

What do students know at the end of their first statistics course? What do they gain in reasoning about statistics from the beginning of the course to the end? Those were the questions that guided an analysis of the data gathered during the Fall 2005 and Spring 2006 class testing of the CAOS 4 test. It was disappointing to see such a small overall increase in correct responses from pretest to posttest, especially when the test was designed (and validated) to measure the most important learning outcomes for students in a non-mathematical, first course in statistics. It was also surprising that for almost all items, there was a noticeable number of students who selected the correct response on the pretest, but chose an incorrect response on the posttest.

The following three broad groups of items emerged from the analyses: (a) items that students seemed to do well both prior to and at the end of their first course, (b) items where they showed the most gains in learning, and (c) items that were more difficult for students to learn. Although less than half of the students were correct on the posttest for all items in the latter category, there was a significant increase from pretest to posttest for almost two thirds of the items in this group. Finally, items were examined that showed an increase in misconceptions about particular concepts. The following sections present a discussion of these results, logically organized by topic areas: data collection and design, descriptive statistics, graphical representations, boxplots, normal distribution, bivariate data, probability, sampling variability, confidence intervals, and tests of significance.

7.1. DATA COLLECTION AND DESIGN

Students did not show significant gains in understanding some important principles of design, namely the purpose of random assignment and that a correlation from an observational study does not allow causal inferences to be drawn. In fact, the percentage of students demonstrating misconceptions increased in terms of believing that random assignment is equivalent to random sampling, or that random assignment reduces sampling error, or that causation can be inferred from correlation.

7.2. DESCRIPTIVE STATISTICS

Students seemed to initially understand the idea of variability of repeated measures. Whereas a small percentage of students made gains in estimating and identifying the histogram with the lowest standard deviation and the graph with the highest standard deviation among a set of histograms, around half of all the students did not demonstrate this ability on the posttest. It seems that some students understood that a graph that is very narrow and clumped in the middle might have less variability, but had different ideas about what more variability might look like (e.g., bumpiness rather than spread from the center). One misconception that increased from pretest to posttest was that a graph with the largest number of different values has the larger standard deviation (spread not considered).

7.3. GRAPHICAL REPRESENTATIONS

Most students seemed to recognize a correct and complete interpretation of a histogram when entering the course, and this did not change after instruction. They did make significant gains in being able to match a histogram to a description of a variable. There was a small increase in the percentage of students who could recognize different graphical representations of the same data, although this was demonstrated by only slightly more than half of the students on the posttest. Only a small percentage of students made gains in understanding that shape, center and spread were represented by a histogram and not a bar graph. One of the most difficult items that showed no significant improvement indicated that students failed to recognize that a distribution with a median larger than the mean is most likely skewed left. Most students were able to make reasonable comparisons of groups using dot plots, and students appeared to gain in their understanding that equal sample sizes are not needed to compare groups

7.4. BOXPLOTS

Students seemed to have many difficulties understanding and interpreting boxplots. A small percentage of students made significant gains in recognizing and interpreting the median in the context of a boxplot. On the posttest, many students seemed to think that the boxplot with the longer lower whisker had a higher percentage of cases below an indicated value or that the boxplot with a longer upper whisker would have a higher percentage of data above the median. Similarly, students did not associate a larger interquartile range with a larger standard deviation, given two boxplots with about the same range. There was no apparent gain in students' understanding that boxplots provide only estimates of percentages at the quartiles.

7.5. NORMAL DISTRIBUTION

Students tended to select responses across various items that showed a normal distribution, suggesting a tendency to select a graph that is like a normal distribution regardless of whether it makes sense to do so within the context of the problem. Presented with an item that reported a median that is noticeably greater than the mean, most students selected a more symmetric, bell-shaped histogram instead of a histogram that is skewed to the left. Many students incorrectly selected a somewhat symmetric, mound-shaped bar graph as a graph that would indicate shape, center and spread, rather than a histogram that was not bell shaped.

7.6. BIVARIATE DATA

Students seemed to do a good job at the beginning of their courses with matching a scatterplot to a verbal description, indicating that they understood how a positive linear relationship was represented on a scatterplot. However, although statistically significant, only a small percentage of students showed gains in recognizing that it is not legitimate to extrapolate using values outside the domain of values for the independent variable when using a regression model. About three fourths of the students did not demonstrate this understanding on the posttest. Of course, it cannot be determined whether the difficulty comes from students not understanding this idea, students not identifying this idea as the focus on the question asked, or the topic not being covered in the course.

7.7. PROBABILITY

The probability topics presented in the CAOS 4 test were quite difficult for students. Students showed no gains from pretest to posttest on items that required identification of correct ratios to use when constructing probabilities from a two-way table, or knowing how to simulate data to find the probability of an outcome.

7.8. SAMPLING VARIABILITY

Students demonstrated difficulty with understanding sampling variability and sampling distributions. There was only a small increase in the percentage of students who demonstrated an understanding that statistics from relatively small samples vary more than statistics from larger samples, an understanding of expected patterns in sampling variability, or an understanding of factors that allow generalization from a sample to a population. Similarly, only a small percentage showed gains on an item that had them select a histogram representing a sampling distribution from a given population for a particular sample size. One of the most difficult items expected them to use sampling error as an appropriate measure when making an informal inference about a sample mean.

7.9. CONFIDENCE INTERVALS

Students did not demonstrate an understanding of confidence intervals. Whereas three fourths of the students recognized a valid interpretation of a confidence interval on the posttest, many of these same students indicated that the invalid statement also applied, as if the two statements had the same interpretation. About two thirds of the students understood that a confidence level does not represent the percentage of population values between the confidence limits. There was an increase in the percentage of students who incorrectly indicated that a confidence level represents the expected percentage of sample values between the confidence limits. The majority of students on the posttest also incorrectly indicated that a confidence level indicated the percentage of all sample means that fall between the confidence limits.

7.10. TESTS OF SIGNIFICANCE

Many students entered the course already recognizing that lack of statistical significance does not mean no effect. Most students indicated on the posttest that a low p-value is required for statistical significance. A small percentage of students made gains in identifying a correct interpretation of a significance test when the null hypothesis is rejected, although almost half did not demonstrate this understanding on the posttest. However, although a little over half of the students recognized a correct interpretation of a p-value, the majority of these students also responded that an incorrect interpretation was valid, indicating that many students hold both types of interpretation without recognizing the contradiction.

8. SUMMARY

The CAOS test provides valuable information on what students appear to learn and understand after completing a college-level, non-mathematical first course in statistics. Across college-level first courses in statistics at a variety of institutions, there were some concepts and abilities that many students demonstrated at the start of a course. These included recognizing a complete description of a distribution and understanding how bivariate relationships are represented in scatterplots. Most students also demonstrated an ability to make reasonable interpretations of some graphic representations by the end of a course. However, the results indicate that many students do not demonstrate a good understanding of much of the content covered by the CAOS 4 test, content that statistics faculty agreed represents important learning outcomes for an introductory statistics course. At the end of their respective courses, students still had difficulty with identifying appropriate types of graphic representations, especially with interpreting boxplots. They also did not demonstrate a good understanding of important design principles, or of important concepts related to probability, sampling variability, and inferential statistics.

It should be noted that all items on the CAOS test were written to require students to think and reason, not to compute, use formulas, or recall definitions, contrary to many instructor-designed exams on which there may be more pretest to posttest gains. However, the CAOS test was purposefully designed to be different from the traditional test written by course instructors. During interviews and on surveys conducted to evaluate the ARTIST project, many instructors communicated that they were quite surprised when they saw their students' scores. They reported that they found the CAOS test results quite illuminating, causing them to reflect on their own teaching in light of the test results. That is one of the most important purposes of the CAOS test, to provide information to statistics instructors to allow them to see if their students are learning to think and reason about statistics, and to promote changes in teaching to better promote these learning goals.

The CAOS test is now available for research and evaluation studies in statistics education. Instructors and researchers can register to use the CAOS test at the ARTIST website (https://app.gen.umn.edu/artist/). Plans are currently underway for the development of a collaborative effort among many institutions to gather large amounts of test data (including CAOS) and instructional data online as a way to promote future research on teaching and learning statistics at the college level. In addition, there is a need to conduct studies that explore particular activities and sequences of activities in helping to improve students' statistical reasoning as they take introductory statistics courses. Given the internal reliability of the CAOS test for students in non-mathematical introductory college statistics courses, and that it has been judged to be a valid measure of important learning outcomes for students enrolled in such courses, we hope that CAOS will facilitate these much needed studies.

ACKNOWLEDGMENT

The Research presented in this manuscript was supported by the National Science Foundation (NSF CCLI ASA-0206571).

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer.

Ben-Zvi, D. (2004). Reasoning about data analysis. In D. Ben-Zvi & J. B. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 121-146). Dordrecht, Netherlands: Kluwer.

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 169-190). Voorburg, The Netherlands: International Statistical Institute.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer.

Cobb, G. (1992). Teaching statistics. In *Heeding the Call for Change: Suggestions for Currricular Action*, *MAA Notes*, *Vol. 22*, 3-33.

delMas, R. & Bart, W. M. (1989). The role of an evaluation exercise in the resolution of misconceptions of probability. *Focus on Learning Problems in Mathematics*, *11*(3), 39-54.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3).
[Online: www.amstat.org/publications/jse/secure/v7n3/delmas.cfm]

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55-82.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_delMas_Liu.pdf]

Garfield, J. (2001). *Evaluating the impact of educational reform in statistics: A survey of introductory statistics courses.* Final Report for NSF Grant REC-9732404.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22-38.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ2 (1).pdf]

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

Garfield, J., delMas, R., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project.* NSF CCLI grant ASA- 0206571.
[Online: https://app.gen.umn.edu/artist/]

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

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309-334.

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41.

Hodgson, T. R. (1996). The effects of hands-on activities on students' understanding of selected statistical concepts. In E. Jakubowski, D. Watkins, & H. Biske (Eds.), *Proceedings of the Eighteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 241–246). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Hogg, R. (1992). Report of workshop on statistics education. In *Heeding the Call for Change: Suggestions for Curricular Action, MAA Notes, Vol. 22*, 34-43.

Howell, D. C. (2002). *Statistical Methods for Psychology (Fifth Edition).* Pacific Grove, CA: Duxbury.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*, 59-98.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3*(1).
[Online: http://www.amstat.org/publications/jse/v3n1/konold.html]

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., Pollatsek, A., Well, A. D., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 34*, 392-414.

Lindman, H., & Edwards, W. (1961). Supplementary report: Unlearning the gambler's fallacy. *Journal of Experimental Psychology*, *62*, 630.

Mathews, D., & Clark, J. (1997, March). Successful students' conceptions of mean, standard deviation, and the Central Limit Theorem. Paper presented at the Midwest Conference on Teaching Statistics, Oshkosh, WI.

McClain, K., Cobb, P., & Gravemeijer, K. (2000). Supporting students' ways of reasoning about data. In M. Burke & F. Curcio (Eds.), *Learning Mathematics for a New Century, 2000 Yearbook*. Reston, VA: National Council of Teachers of Mathematics.

Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal, 1*(2), 22-37.
[Online: http://fehps.une.edu.au/serj]

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*, 123-137.

Pedhazur, E. J., and Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Erlbaum.

Pollatsek, A., Konold, C., Well, A., and Lima, S. (1984). Beliefs underlying random sampling. *Memory and Cognition, 12*(4), 395-401.

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

Reed-Rhoads, T., Murphy, T. J., & Terry, R. (2006). *The Statistics Concept Inventory (SCI)*.
[Online: http://coecs.ou.edu/sci/]

Rubin, A., Bruce, B., & Tenney, Y. (1990, August). Learning about sampling: Trouble at the core of statistics. Paper presented at the Third International Conference on Teaching Statistics, Dunedin, New Zealand.

Scheaffer, R. (1997). Discussion. *International Statistical Review*, *65*, 156-158.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Erlbaum.

Saldanha, L. A., & Thompson, P. W. (2001). Students' reasoning about sampling distributions and statistical inference. In R. Speiser & C. Maher (Eds.), *Proceedings of The Twenty-Third Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 449-454), Snowbird, Utah. Columbus, Ohio: ERIC Clearinghouse.

Shaughnessy, M. (1977). Misconceptions of probability: An experiment with a small-group, activity-based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics, 8*, 295-316.

Shaughnessy, M. (1992). Research in probability and statistics: Reflections and directions. In A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 465-494). New York: MacMillan Publishing Company

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999, April). School mathematics students' acknowledgment of statistical variation. For the NCTM Research Presession Symposium: *There's More to Life than Centers*. Paper presented at the 77th Annual National Council of Teachers of Mathematics (NCTM) Conference, San Francisco, CA.

ROBERT DELMAS
157 Education Sciences Building
56 East River Road
University of Minnesota
Minneapolis, MN 55455-0364
USA

APPENDIX A: PERCENT OF STUDENTS WITH ITEM RESPONSE PATTERNS FOR SELECTED CAOS ITEMS

			Item Response Pattern[a]			
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post
1	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.	760	8.6	17.9	20.4	53.2
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram).	759	26.0	17.8	28.6	27.7
3	Ability to visualize and match a histogram to a description of a variable (neg. skewed distribution for scores on an easy quiz).	760	20.8	6.1	22.5	50.7
4	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).	757	26.6	10.3	25.5	37.6
5	Ability to visualize and match a histogram to a description of a variable (uniform distribution for the last digit of phone numbers sampled from a phone book).	758	23.0	5.9	21.1	50.0
6	Understanding to properly describe the distribution of a quantitative variable, need a graph like a histogram that places the variable along the horizontal axis and frequency along the vertical axis.	754	68.0	6.8	16.8	8.4
7	Understanding of the purpose of randomization in an experiment.	754	81.2	6.5	10.3	2.0
8	Ability to determine which of two boxplots represents a larger standard deviation.	755	21.3	19.5	24.0	35.2
9	Understanding that boxplots do not provide estimates for percentages of data above or below values except for the quartiles.	751	59.7	13.7	17.0	9.6

[a] Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest.

			Item Response Pattern[a]			
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post
10	Understanding of the interpretation of a median in the context of boxplots.	754	62.3	9.4	18.0	10.2
11	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities.	756	3.8	7.9	8.2	80.0
12	Ability to compare groups by comparing differences in averages.	753	4.8	9.4	10.0	75.8
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.	752	15.4	11.0	22.7	50.8
14	Ability to correctly estimate and compare standard deviations for different histograms. Understands lowest standard deviation would be for a graph with the least spread (typically) away from the center.	746	38.6	9.7	27.1	24.7
15	Ability to correctly estimate standard deviations for different histograms. Understands highest standard deviation would be for a graph with the most spread (typically) away from the center.	747	37.1	16.1	24.6	22.2
16	Understanding that statistics from small samples vary more than statistics from large samples.	747	60.2	7.9	17.0	14.9
17	Understanding of expected patterns in sampling variability.	746	37.3	12.5	20.0	30.3
18	Understanding of the meaning of variability in the context of repeated measurements and in a context where small variability is desired.	746	7.6	11.8	11.8	68.8
19	Understanding that low p-values are desirable in research studies.	730	21.1	10.4	32.7	35.8
20	Ability to match a scatterplot to a verbal description of a bivariate relationship.	748	1.9	5.6	7.6	84.9

			Item Response Pattern[a]			
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).	749	7.1	9.2	19.4	64.4
22	Understanding that correlation does not imply causation.	743	27.5	19.9	17.9	34.7
23	Understanding that no statistical significance does not guarantee that there is no effect.	735	17.6	18.1	19.3	45.0
24	Understanding that an experimental design with random assignment supports causal inference.	731	20.1	20.4	21.3	38.2
25	Ability to recognize a correct interpretation of a p-value.	712	23.5	22.1	29.8	24.7
26	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is not effective).	719	19.5	22.0	27.4	31.2
27	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is effective).	717	28.5	18.8	29.3	23.4
28	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits)	729	33.5	23.3	18.1	25.1
29	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).	725	24.4	8.0	43.0	24.6
30	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits)	723	38.9	16.9	29.7	14.5
31	Ability to correctly interpret a confidence interval.	720	16.0	9.7	36.9	37.4
32	Understanding of how sampling error is used to make an informal inference about a sample mean.	718	70.2	12.7	13.0	4.2

[a] Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest.

			Item Response Pattern[a]			
Item	Measured Learning Outcome	*n*	Incorrect	Increase	Decrease	Pre & Post
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left.	730	36.6	23.7	21.9	17.8
34	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.	724	19.1	15.7	25.7	39.5
35	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.	719	39.4	16.4	26.1	18.1
36	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.	719	25.7	21.3	21.6	31.4
37	Understanding of how to simulate data to find the probability of an observed value.	722	67.3	13.2	12.3	7.2
38	Understanding of the factors that allow a sample of data to be generalized to the population.	715	50.5	11.6	23.5	14.4
39	Understanding of when it is not wise to extrapolate using a regression model.	710	63.9	11.5	18.2	6.3
40	Understanding of the logic of a significance test when the null hypothesis is rejected.	716	31.6	16.5	26.5	25.4

[a] Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest.

APPENDIX B: PERCENT OF STUDENTS WITH ITEM RESPONSE PATTERNS FOR CAOS ITEMS ASSESSING MISUNDERSTANDING AND MISCONCEPTIONS

			Item Response Pattern[a]			
Item	Misconception or Misunderstanding	*n*	Neither	Decrease	Increase	Pre & Post
6	A bell-shaped bar graph to represent the distribution for a quantitative variable.	754	31.6	15.6	25.5	27.3
7	Random assignment is confused with random sampling or thinks that random assignment reduces sampling error.	754	36.5	14.3	27.3	21.9
15	When comparing histograms, the graph with the largest number of different values has the larger standard deviation (spread not considered).	747	52.9	14.1	20.6	12.4
22	Causation can be inferred from correlation.	743	50.9	13.2	22.1	13.9
36	Grand totals are used to calculate conditional probabilities.	719	51.3	15.3	23.5	9.9
40	Rejecting the null hypothesis means that the null hypothesis is definitely false.	716	50.4	17.2	22.9	9.5

[a] Neither = did not select the response on either the pretest or the posttest; Decrease = response selected on pretest, but not on the posttest; Increase = response not selected on the pretest, selected on the posttest; Pre & Post = response selected on both the pretest and posttest.

EVALUATION OF DISTANCE LEARNING IN AN "INTRODUCTION TO BIOSTATISTICS" CLASS: A CASE STUDY

SCOTT R. EVANS
Harvard University Extension School
evans@sdac.harvard.edu

RUI WANG
Harvard University Extension School
rwang@hsph.harvard.edu

TZU-MIN YEH
Harvard University Extension School
tyeh@hsph.harvard.edu

JEFF ANDERSON
Harvard University Extension School
janderson@sdac.harvard.edu

RAMMY HAIJA
Harvard University Extension School
rammyhaija@hotmail.com

PAUL MADOC MCBRATNEY-OWEN
Harvard University Extension School
mcbratn2@fas.harvard.edu

LYNNE PEEPLES
Harvard University Extension School
lpeeples@hsph.harvard.edu

SUBIR SINHA
Harvard University Extension School
ssinha@sdac.harvard.edu

VANESSA XANTHAKIS
Harvard University Extension School
vxanthakis@mclean.harvard.edu

NATASA RAJICIC
Harvard University Extension School
nrajicic@hsph.harvard.edu

JIAMENG ZHANG
Harvard University Extension School
jiamengz@hsph.harvard.edu

Statistics Education Research Journal, 6(2), 59-77, http://www.stat.auckland.ac.nz/serj

ABSTRACT

Biostatistics is not universally available in colleges/universities and is thus an attractive course to offer via distance education. However, evaluation of the impact of distance education on course enrollment and student success is lacking. We evaluated an "Introduction to Biostatistics" course at Harvard University that offered the distance option (Spring 2005).We assessed the effect on course enrollment and compared the grades of traditional students with non-traditional students, as well as with historical traditional students (Fall 2004). We further compared course evaluations from the inaugural semester with the distance option to evaluations from the prior semester. No evidence of dissimilarities was noted with respect to overall course grade averages or course evaluations.

Keywords: *Statistics education research; Biostatistics; Distance education*

1. INTRODUCTION

Time and geographical constraints make distance education a convenient and appealing option for many students. Wegman and Solka (1999) note that distance education is a wave of the future, particularly with a recent increased emphasis on re-educating a work force engaged in a lifetime of learning. Gilmour (2002) notes that distance education courses are likely to rise with expanding access to the Internet. Increases in the number of classes offered with a distance option, and in the number of distance students, have been observed at many colleges and universities. According to a recent national survey by the Sloan Consortium (Allen & Seaman, 2005), an online education group, at least 2.3 million people took an online course in 2004 and two-thirds of institutions offering traditional undergraduate-level courses also offer online courses, and similarly for graduate-level course offerings. The Harvard University Extension School initiated a distance education program with a single course enrolling four distance students in 1997-1998. This program has grown to 55 courses offered with a distance option in 2004-2005 and 75 courses being offered in 2005-2006.

Biostatistics courses are not universally offered at colleges and universities (e.g., institutions without graduate programs in public health and/or medicine frequently do not offer courses in biostatistics). However, the demand for biostatistics courses is high due to the increasing needs of medical students, pre-medical students, public health students, as well as employees in the pharmaceutical and biotechnology industries, research hospitals, government, and academia. Thus, an "Introduction to Biostatistics" course is a particularly attractive course to offer via distance education. However, careful evaluation of the effectiveness of such a course on student learning and student evaluation is lacking.

An "Introduction to Biostatistics" course offered at Harvard University was evaluated. Grades of traditional students were compared to non-traditional students enrolled in a course open to both traditional and non-traditional students in the spring of 2005. Traditional students in the course were further compared to historical traditional students (Fall 2004) to investigate whether the distance version of the course affected the traditional students in terms of grades and evaluation scores. In addition, course evaluations for the semester in which the course was offered with a distance option (Spring 2005) and for the semster in which the course was offered only traditionally (Fall 2004) were compared.

We summarize the results of this research in the following sections. We describe the methods of delivery for distance education, discuss pros and cons associated with distance education, and summarize prior evaluations of distance learning in statistics and biostatistics in Section 2. In Section 3, we describe the objectives and methods of our study. In Section 4, we describe the results of our study, and then conclude with a discussion of limitations, recommendations, and future research in Section 5.

2. DISTANCE EDUCATION

Although distance education has become common, its use is controversial and research suggests that its effectiveness is variable and inconsistent (Rooney et al., 2006; Rivera, McAlister, Khris, & Margaret, 2002; Li, 2002). Many research studies, whether comparison or case studies, have shown that distance learning is as favorable as classroom learning and that distance students are satisfied, and have similar grades or test results, compared to traditional students (Phipps & Merisotis, 1999; Johnson, Aragon, Shai, & Palma-Rivas, 1999; Russell, 1999; Merisotis & Phipps, 1999; Bourne, McMaster, Rieger, & Campbell, 1997; Gagne & Shepherd, 2001). A meta-analysis by Allen, Bourhis, Burrell, and Mabry (2002) demonstrated little difference in satisfaction levels of students between the traditional and distance educational formats. However, other research suggests that distance students may not be learning the material as well as those enrolled in traditional classes (Clow, 1999), and that use of the Internet merely to post materials and return homework may result in poorer learning than in traditional classes (Hiltz, Coppola, Rotter, & Turoff, 2001).

An e-mail to the group list of the American Statistical Association’s (ASA) Section on Teaching Statistics in the Health Sciences (TSHS) requesting guidance from experienced distance education instructors in preparation for this “Introduction to Biostatistics” course generated numerous responses, with varying degrees of support. “Distance Education: How’s it working?” was a panel discussion at Joint Statistical Meetings (JSM), 2004, and the aforementioned e-mail discussions resulted in an invited session, “Distance Learning in the Health Sciences” at JSM 2005, sponsored by the Section for TSHS. The Section on TSHS also sponsored a roundtable luncheon, “Distance Education in Biostatistics” at JSM 2005, and the Boston Chapter of ASA hosted a mini-conference on Distance Learning, “Distance Education, The Way of the Future?” Open discussions at these meetings suggested that research regarding the effectiveness of student learning in distance courses is needed. Many observations regarding the advantages and disadvantages of such courses were discussed (and are outlined in Section 2.2).

2.1. METHODS OF DELIVERY

Distance education is a very broad term that encompasses several methods of delivery. Although distance education is generally thought of as a recent development using state-of-the-art technology, correspondence courses are a form of non-electronic distance education that have been in use for many years. Typically, students register for the course and then receive a course packet including a syllabus, reading instructions, and homework problems. Completed homework assignments are sent by the student to the corresponding instructor by postal mail or fax. The instructor corrects the assignment, provides comments, and returns the graded homework. These courses are often self-paced and do not necessarily adhere to a strict semester schedule.

Online delivery has become a widely-used method due to its rapid delivery and response time. It may take the form of a lecture post in which the instructor posts a documented lecture and instructions online. Students read the lecture and then post questions for response from the instructor and other students. It may also take the form of a lecture feed where the instructor may lecture to an attending group of students while being recorded, and then the lecture is made available for access. Videos may be made available at specified remote sites where groups of students congregate to view the video together at specified times, or online where students can individually view videos with use of the Internet. Online delivery may also be synchronized such that distance students log-on to a course website at the scheduled time of the course to view the lecture live, and can interactively communicate with the instructor while the lecture is being taught. Such courses are usually offered on a standard school semester schedule. Thus students must keep up with the lectures or risk falling behind in the course. Live chat room sessions often serve as an additional communication supplement.

A number of software tools are available to facilitate or supplement distance education, such as WebCT (Web Course Tools), Blackboard®, NetMeeting, Centra®, and Elluminate *Live!*®, and Moodle. WebCT, developed for academic use by Dr. Murray Goldberg and a group of colleagues at the University of British Columbia, and Blackboard provide many features such as announcements, bulletin boards, chat rooms, virtual classrooms, group forums, private e-mail, searching, quizzes, surveys, student home pages, glossary, syllabus, contact information, digital drop box, and gradebook (Kendall, 2001; Wernet, Olliges, & Delicath, 2000). [In February 2006, Blackboard Inc. completed a merger with WebCT Inc.]. Microsoft® NetMeeting is a Windows-based application that can be used for synchronous activities. It allows synchronous chatting, application sharing, and file sharing. Tools built into NetMeeting include whiteboard, chat, file transfer, program sharing, and remote desktop sharing. Centra® is a web-based software application that enables real-time communication collaboration and learning with features of virtual classes, web seminars, and eMeetings. Elluminate Live!® creates a real-time virtual classroom environment for distance learning and collaboration, with its many components such as two-way audio, live video, shared whiteboards, instant messaging, application sharing, and breakout rooms. Moodle, a course management system (CMS) created by Martin Dougiamas at Curtin University, Australia, is a free, open source software package designed to help educators create effective online learning communities. It has many of the above features in addition to blogs, wikis, database activities, peer assessment, and multi-language support, and can readily be extended by creating plugins for specific new functionality.

2.2. ADVANTAGES AND DISADVANTAGES

Several advantages and disadvantages associated with distance education have been identified. Many generally apply to various courses; some are of increased importance in statistics and biostatistics.

Advantages of distance education include the following:

1. There may be no alternative for many students. For example, many undergraduate colleges do not offer courses in biostatistics. Distance courses can provide the opportunity for students from such colleges to take biostatistics courses. More generally, distance education offers flexibility and convenience, allowing geographically isolated students, and students with conflicting time commitments, to continue their education. A distance option also provides students with the opportunity to take two courses that are offered at the same time. Further, the distance option may allow students the opportunity to take a course from a prominent expert in the field.

2. If the distance option includes video, then students have the ability to watch the video as many times as desired. If the student finds the course material difficult, or the student misses a class, then the video may provide a useful learning tool. Stephenson (2001) notes that distance students may take a break when tired by stopping the video, whereas students in a traditional course do not have the same luxury. Because the material is always available, students have control over the pace of learning. This can be particularly important for courses in statistics or biostatistics in which students often struggle with statistical concepts and notation.
3. Distance education may sharpen teaching skills, as both diligent preparation and careful delivery are required to teach a distance education course.
4. Successful programs of teaching statistics via distance have been documented. Speed and Hardin (2001) present the results of developing technology mediated instructional material (TMIM) for graduate level statistics courses presented to students at local and distance sites. Improvements are also possible as teachers learn how to teach, and students learn how to learn, using distance education resources.
5. Online learning may be cost effective because internet-based courses can be made available to an almost infinite number of students (Katz & Yablon, 2003b). However, cost effectiveness of using online technologies in distance education is still uncertain (Phelps, Wells, Ashworth, & Hahn, 1991), as human capital and the costs of conversion are expenses that can easily be underestimated (Ng, 2000). Carr (2001) argues that only in large courses, with many sections, would cost savings be possible. The startup costs, maintenance costs, and personnel costs should be factored in to arrive at a true cost of a distance-learning program (Valentine, 2002). Hillstock (2005) further argues that distance learning as a way to save money is a misconception. Notably, some responses to the TSHS group list indicated instructors' fear of job insecurity due to the availability of distance courses.
6. Instructors who have recorded lectures (e.g., videotapes) from prior semesters may use these recordings in future classes. This may be an attractive option when an instructor is sick or traveling, or if an instructor feels that he or she explained a topic particularly well in a specific recorded lecture.

Several disadvantages of distance education have been observed or theorized including the following:

1. Distance education may allow students to become lazy, using the online component as a crutch. A student may feel that missing a class can be justified with the availability of online videos that can be viewed at a later time. As a result, students may fall behind more frequently. This can be particularly problematic in statistics and biostatistics courses where comprehension of later concepts is conditional upon comprehension of earlier topics. Students may not be able to recover in such courses.
2. In many forms of distance education, there is no live communication between the distance student and the instructor. Many instructors feel that the face-to-face contact and the student/teacher interaction are critical to learning. Lack of this personal element also makes it more difficult for instructors to stimulate, motivate, or excite students. Students can feel a sense of isolation

and a lack of support. Students in statistics and biostatistics may need feedback on difficult concepts or computing issues.

3. Non-verbal communication such as body language and facial expressions, as well as verbal cues, cannot be conveyed with distance education. Often instructors can recognize whether a class is "getting it" by facial expressions of the students in traditional courses. The "reading" of such students is not possible in distance courses.
4. Group projects may be more difficult for distance students as it may not be possible for students to meet face-to-face. Therefore, students may not be able to learn from each other as effectively. Furthermore, there may be less satisfaction without group interaction.
5. "Asynchronous" distance students are at a time disadvantage as they do not have the opportunity to ask direct questions in a timely manner (if at all). There are typically delays in both video access and receiving graded homework assignments.
6. Technology problems may create a disadvantage for distance students. Students must often solve computing hardware or software issues on their own. Foster (2003) notes that the mathematical notation involved in biostatistics and statistics may also create technological obstacles with respect to software. Such technology issues make it difficult to provide students with equal access to course materials.
7. Distance courses may become "watered-down" because instructors (intentionally or subconsciously) may be sympathetic to the additional complications involved with distance education. This could jeopardize education quality.
8. Teaching distance education courses requires more diligence and preparation than traditional courses. Lectures need to be self-explanatory (Bruce, Bond, & Jones, 2002). Instructors often underestimate the time and resources needed for development of course materials. The volume of e-mail created from distance students may increase dramatically for the instructor and teaching assistants.
9. Institutions are often not well prepared for offering distance courses, and methods of delivery may not be sufficient for effective learning. Due to budget limitations, instructors may not have sufficient school support (e.g., funding) and access to resources for dealing with the issues created by distance education.
10. Distance education may create more problems with cheating and academic honesty. It may be more difficult to assess whether students have completed their own work.

2.3. EVALUATION OF DISTANCE LEARNING IN BIOSTATISTICS AND STATISTICS

A few evaluations of distance learning in statistics have been published. Katz and Yablon (2003a), through use of an external control group, found that students in an internet-based "Introduction to Statistics" course achieved similar grades, were characterized by a higher locus of control, and had higher motivation and satisfaction than students in a lecture-based course. Traditional students had higher levels of self-esteem, however. Katz and Yablon (2003b) further suggest that as students gain more experience with distance learning, students become more at ease and develop positive attitudes toward this educational format.

Harrington (1999) compared grades of incoming Masters of Social Work (MSW) students in traditional versus distance statistics courses. Students with high prior grade point averages (GPAs) performed similarly regardless of registration status (distance or traditional). However, students with low GPAs performed better than traditional students.

Stephenson (2001) compared grades of distance and traditional students in an "Applied Statistics for Industry" course and found that traditional students had slightly higher grade point averages, but noted that the differences may be due to random variation.

McGready (2006) conducted a non-randomized study comparing the exam grades of on-campus and on-line students in a "Statistical Reasoning in Public Health" course at Johns Hopkins University. The study consisted of separate independent lectures (but using the same slide set) for the distance and traditional versions of the course. The distance version of the course utilized streaming audio and synced slides. No statistically significant between-group differences were identified.

3. METHOD

3.1. INTRODUCTION TO BIOSTATISTICS

"Introduction to Biostatistics" (STAT E-102) at the Harvard University Extension School was an introduction to statistical methods course used in the public health, biological, and medical sciences. Topics included descriptive statistics, performance characteristics of diagnostic tests, graphical methods, estimation, hypothesis testing, p-values, confidence intervals, correlation, linear regression, and clinical trials. The course did not require any formal pre-requisites and was offered for graduate or undergraduate credit.

The Instructor for the course had a PhD (Biostatistics) with a primary appointment in the Harvard School of Public Health. Seven experienced Teaching Assistants (TAs) were involved with the course. Six of the TAs had master's degrees (4 Biostatistics/Statistics, 1 Epidemiology, 1 Mathematics) and one had a PhD (Biostatistics).

The course had one two-hour lecture per week (also available via video) with five optional TA help sessions offered throughout the week. Six homework assignments were collected during the semester. Graduate students also completed two projects. Grades for graduate students were determined by the midterm exam (30%), the final exam (30%), homework (20%), and the projects (20%). Grades for undergraduates were determined by the midterm exam (35%), the final exam (40%), and homework (25%).

Students were required to learn a software package of their choice. Support was provided for STATA, with weekly TA sessions (for local students only) and handouts. The course text was *Principles of Biostatistics* by Pagano and Gauvreau (2000).

The course had an extensive and active website. All course materials were posted on the website, including the syllabus, lecture notes, lecture videos, homework assignments and solutions, project assignments and solutions, computing (STATA) handouts, course announcements, and a history of questions and answers. A course e-mail distribution list provided the opportunity for students to ask questions of the TAs and communicate with other students, and further provided the TAs with the opportunity to make announcements regarding homework assignments and exams.

3.2. VIDEO DELIVERY AND DISTANCE METHODS

Lectures available on the course website used streaming video technology (i.e., video that is sent to the user as it is viewed) along with standard Internet browser software. The video window appeared on the left side of the screen, the controls below it adjusted the video, and the supplementary course materials were displayed on the right side of the screen. The advantage of streaming video is that, like TV or radio, students receive the images and audio just before they see and hear them. This is much quicker than waiting for the entire video file to download before viewing it, as is the case with static images on the Web. The disadvantage is that in order to decrease file size and allow for a steady stream of data, the video must be compressed, shrinking the image from full size.

Lectures were typically available within 48 hours after they were presented on campus. Recorded lectures were available only to registered students; lectures were password protected after the second week of class.

Students downloaded and installed current versions of one of the supported video players before attempting to view the lectures. Students were responsible for ensuring that they had the necessary computer hardware and software, including course-specific software needed to complete course assignments. Harvard University did not provide equipment or software. No toll-free dial-in access was available.

Lecture notes were primarily in PowerPoint, allowing use of an automated time-stamp file as an aid in producing the video and in synchronizing the slides with the video. However, a manual time-keeper also recorded the timing of the slides, as other non-PowerPoint slides were also utilized.

Students living in the six-state New England area (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont) were required to take all classroom examinations on campus as scheduled. Distance education students outside the New England area arranged to take their exams off site and submitted a completed distance education proctored exam form to Academic Services at least two weeks in advance.

Distance students could submit homework via e-mail, fax, or postal mail through a centralized address at the Harvard Extension School.

3.3. OBJECTIVES

“Introduction to Biostatistics” was taught at the Harvard Extension School in each of the Fall 2004 and Spring 2005 semesters. The course was taught traditionally (without a distance option, e.g., without video) in Fall 2004. However, in Spring 2005, the course was offered with a distance option that included video access for all students in the course. Students could register for the course (hereafter termed registration status) as distance students, traditional students, or hybrid students (i.e., students that attend some live lectures of their choice in person but use the video for other lectures). Non-traditional students consisted of hybrid and distance students combined.

This research has several objectives:

1. To examine the effect on enrollment of adding a distance option to an Introduction to Biostatistics course.
2. To evaluate the “distance effect” by a comparison of grades of non-traditional students vs. traditional students in Spring 2005. The traditional students in this course provide a unique “internal” control group that can be utilized to evaluate the distance effect. Semester-to-semester variation is eliminated by the use of this internal control group.

3. To evaluate the “video effect” by a comparison of grades between traditional students in a course with a distance option that included access to lecture videos (Spring 2005) and traditional students in the same course offered without a distance option, and thus without video access (Fall 2004). Variation due to registration status is eliminated by using only traditional students in the control group.
4. To compare course evaluations for the two semesters. Because course evaluations are anonymous, the registration status of a responder cannot be determined. Thus evaluation of a distance effect, or a video effect, with respect to evaluations is not possible.

3.4. RESEARCH QUESTIONNAIRE AND SCHOOL SUPPORT

An “Information for Research Subjects and Research Questionnaire” was provided to students enrolled in each class. The questionnaire provided the students with a description of the study, information regarding confidentiality, potential risks and benefits, time commitment, and whom to contact with questions. The Information for Research Subjects and Research Questionnaire was reviewed and approved (judged exempt) by the Human Subjects Committee at the Harvard School of Public Health and by the Research Review Committee at Academic Services at the Harvard Extension School. As per the request of these committees, the instructor did not view the questionnaire until grades were finalized. A student signature was also required to include the student’s grade in this evaluation. Students were treated equally regardless of their decision to participate.

The one-page Research Questionnaire collected data such as demographics, educational level, place of employment, whether the student had a prior statistics or biostatistics course, whether the student was a mathematics or statistics major, and whether the course was a requirement.

The Extension School provided funding for two Faculty Aides (i.e., students that required research experience in order to complete their program) to assist with this research project.

3.5. STATISTICAL CONSIDERATIONS AND METHODS

Descriptive statistics are used to describe the study sample. In general, categorical variables are summarized with counts and percentages. Continuous variables are summarized by displaying descriptive statistics (*n*, mean, standard deviation, and median). A bar graph is used to display student enrollment over time. Between-group comparisons are performed using Wilcoxon Rank Sum tests for continuous baseline variables and grades, Fisher’s exact tests for categorical baseline variables, and the mean score test (Stokes, Davis, & Koch, 1995) was used to compare the evaluations. Confidence intervals (CIs) using the *t* distribution are used to estimate evaluation rating differences. Exact CIs are used to estimate between-group difference in grades.

All significance testing is performed at the 0.05 level and all reported p-values are two-sided. There is no adjustment for multiple testing, therefore results should be interpreted with caution.

4. RESULTS

4.1. ENROLLMENT

A substantial enrollment increase (100%) was noted when the distance option was offered. Figure 1 displays enrollment for the past six semesters (all taught by the same instructor) for this course. It is notable that this increase occurred without targeted advertising of the distance option. The course enrollment consisted of only 10% distance students but 39% hybrid students. This suggests that the freedom and flexibility of lecture or video availability offered by the hybrid option was particularly attractive to students.

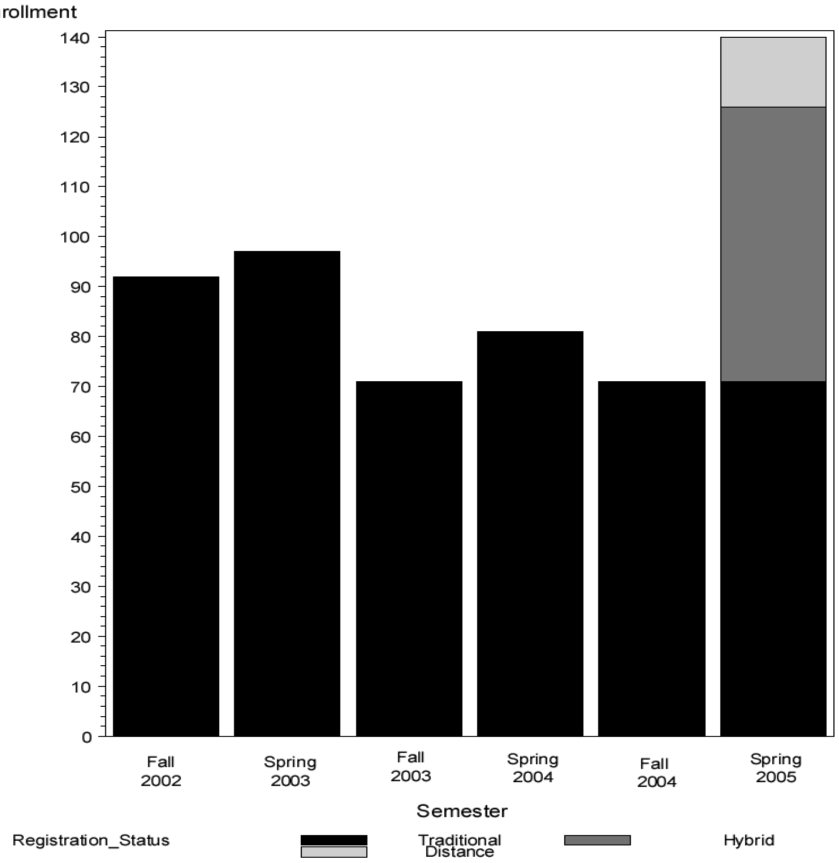


Figure 1. Student enrollment over time

4.2. DEMOGRAPHICS AND BASELINE CHARACTERISTICS

Table 1 displays demographic and baseline characteristics of students in the courses by semester. These data were collected via a questionnaire to assess similarity of the comparison groups at baseline (i.e., upon entering the course). Summaries are provided only for students who signed the Research Questionnaire. No significant differences between comparison groups are noted with respect to age, gender, race, education, and place of work. Further, the proportions of students that were graduate students, had a prior statistics course, were mathematics or statistics majors, and for which this was a required course were also not dissimilar.

Table 1. Summary of Student Demographics and Baseline Characteristics

Variable	Fall 2004 (*n*=52)	Spring 2005				p-value[a]	p-value[b]
		Trad. (*n*=43)	Hybrid (*n*=28)	Distance (*n*=8)	Total (*n*=79)		
Age (Mean/SD/Median)	31/9/28	29/8/26	29/7/27	35/10/39	30/8/26	.508	.283
Gender (*n*, %)						.468	.503
Male	17 (33%)	11 (26%)	9 (32%)	3 (37%)	23 (29%)		
Female	35 (67%)	32 (74%)	19 (68%)	5 (63%)	56 (71%)		
Race (*n*, %)						.490	.726
White	37 (71%)	34 (79%)	18 (64%)	6 (75%)	58 (73%)		
Black	1 (2%)	2 (5%)	2 (7%)	0 (0%)	4 (5%)		
Hispanic	3 (6%)	1 (2%)	0 (0%)	0 (0%)	1 (1%)		
Asian	10 (19%)	5 (12%)	6 (21%)	1 (13%)	12 (15%)		
Native American	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)		
Other	1 (2%)	1 (2%)	2 (7%)	1 (13%)	4 (5%)		
Education[c] (*n*, %)							
Bachelor’s	50 (96%)	37 (86%)	24 (86%)	6 (75%)	67 (85%)	.763	.135
Master’s	11 (21%)	8 (19%)	5 (18%)	0 (0%)	13 (16%)	.762	.802
PhD	5 (10%)	1 (2%)	1 (4%)	1 (13%)	3 (4%)	.589	.217
MD	3 (6%)	1 (2%)	1 (4%)	1 (13%)	3 (4%)	.589	.624
Prior Stat Course	31 (60%)	24 (56%)	11 (39%)	5 (63%)	40 (51%)	.370	.835
Math/Stat Major	2 (4%)	1 (2%)	1 (4%)	0 (0%)	2 (3%)	>.995	>.995
Required Class	14 (17%)	16 (37%)	8 (29%)	3 (38%)	27 (34%)	.636	.376
Workplace (*n*, %)						.907	.303
Student	8 (15%)	3 (7%)	2 (7%)	1 (13%)	6 (8%)		
Academic	6 (12%)	7 (16%)	10 (36%)	0 (0%)	11 (22%)		
Hospital	10 (19%)	14 (33%)	6 (21%)	4 (50%)	24 (30%)		
Pharmaceutical	6 (12%)	1 (2%)	1 (4%)	0 (0%)	2 (3%)		
Biotechnology	8 (15%)	4 (9%)	3 (11%)	1 (13%)	8 (10%)		
Government	3 (6%)	2 (5%)	0 (0%)	1 (13%)	3 (4%)		
Other	11 (21%)	12 (30%)	6 (21%)	1 (13%)	19 (24%)		
Credit Level (*n*, %)						>.995	.538
Graduate	25 (48%)	24 (56%)	17 (61%)	4 (50%)	45 (57%)		
Undergraduate	27 (52%)	19 (44%)	11 (39%)	4 (50%)	34 (43%)		

[a]Traditional students vs. non-traditional students in Spring 2005. [b]Traditional students in Spring 2005 vs. traditional students in Fall 2004. [c]Responses are not mutually exclusive.

Table 2. Summary of Student Grades (Spring 2005 Traditional vs. Non-Traditional)

Variable (Mean/SD/Median)	Total	Traditional	Hybrid	Distance	p-value[a]	Difference (95% CI)[b]	Difference (95% CI)[c]	Difference (95% CI)[d]
Total (*n*)	79	43	28	8				
Overall grade	88/11/91	89/12/91	89/9/91	85/13/87	0.82	(-4.2, 5.3)	(-13.3, 6.3)	(-4.5, 5.4)
Exam average	82/14/85	82/15/83	83/12/85	78/19/83	0.83	(-5.0, 7.5)	(-17.5, 7.5)	(-7.5, 7.5)
Homework average	92/8/94	92/8/95	93/7/95	91/5/93	0.66	(-1.7, 2.5)	(-5.5, 2.3)	(-2.3, 2.3)
Project average	94/7/96[e]	94/7/97[f]	94/7/96[g]	89/8/89[h]	0.47	(-1.5, 4.0)	(-14.5, 1.5)	(-3.5, 2.0)

[a] Wilcoxon rank sum test comparing Spring 2005 traditional vs. non-traditional students (hybrid and distance combined), stratified by credit level.

[b] Exact CI for Spring 2005 traditional vs. non-traditional students.

[c] Exact CI for Spring 2005 traditional vs. distance students.

[d] Exact CI for Spring 2005 traditional vs. hybrid students (Hodges & Lehmann, 1963; Lehmann, 1975).

[e] *n* = 45. [f] *n* = 24. [g] *n* = 17.· [8] *n* = 4.

4.3. GRADES

Evaluation of the Distance Effect We compared the grades (overall course average, exam average, and homework average) of the traditional students to the non-traditional students in Spring 2005 in order to evaluate the distance effect, stratifying by credit level (graduate or undergraduate). The independent variable was registration status, and the dependent variable was grade. Results are displayed in Table 2.

No statistically significant differences between traditional versus non-traditional students (Spring 2005) were noted with respect to overall course average, exam average (i.e., equally weighted average of the midterm and final exams), homework average, or project average. CI estimates (95%) for the difference between traditional and non-traditional students indicate that group differences for the overall median grade may be as large as 5 points in either direction.

Evaluation of the Video Effect We further compared the traditional students in Spring 2005 to the traditional students from Fall 2004 to evaluate the video effect, stratifying by credit level (graduate or undergraduate). The independent variable was semester and the dependent variable was grade. Table 3 displays the results.

Statistically significant differences with respect to exam average were noted between traditional students in Spring 2005 and Fall 2004, with the students in Spring 2005 performing worse. Confidence interval estimates indicate that median group differences for the overall grade may be as large 12.5 points. This is believed to be caused by (unintentional) semester-to-semester variation in exam difficulty. However, one cannot rule out a detrimental effect of the distance option on traditional students, potentially due to camera distractions in class or a diversion of the instructors' attention to address distance student needs. No statistically significant differences were noted with respect to overall course average, homework average, or project average.

Table 3. Summary of Student Grades (Fall 2004 Traditional vs. Spring 2005 Traditional)

Variable (Mean/SD/Median)	Fall 2004 Total	Spring 2005 Traditional	p-value[a]	Difference (95% CI)[b]
Total (*n*)	52	43		
Overall grade	92/10/95	89/12/91	0.11	(-7.6, 0.8)
Exam average	91/11/95	82/15/83	<0.01	(-12.5, -2.5)
Homework average	90/10/94	92/8/95	0.51	(-1.3, 2.7)
Project average	92/6/93[c]	94/7/97[d]	0..08	(0.0, 6.0)

[a]Wilcoxon rank sum test comparing Fall 2004 vs. Spring 2005 traditional students, stratified by credit level. [b]Exact CI for Fall 2004 vs. Spring 2005 traditional students (Hodges & Lehmann, 1963; Lehmann, 1975). [c]*n*=25. [d]*n*=24.

4.4. COURSE EVALUATIONS

Course evaluations were voluntary, anonymous, and were not reviewed by the instructor until grades had been finalized and submitted. Thus, we were unable to distinguish evaluations by registration status. However, we compared the evaluations from Spring 2005 to the evaluations from Fall 2004 to examine if the addition of the distance option affects course evaluations. In particular, we compared responses to "rate the course overall" and "rate the instructor overall." Each was rated on a scale from 1 (poor) to 5 (very good). The dependent variables were evaluation scores for the course and the instructor, and the independent variable was semester. Results are displayed in Table 4. No significant differences between semesters were noted with respect to the overall course rating or the overall instructor rating. The CIs for the mean difference of the evaluations between semesters (Spring 2005 minus Fall 2004) provide ranges of potential differences between evaluations for the overall course and the overall instructor ratings.

Table 4. Course Evaluation Summary

Overall Rating	Fall 2004	Spring 2005	p-value[a]	95% CI[b]
Course				
n	50	58		
Mean	3.94	4.09		(-0.18, 0.47)
S.D.	0.87	0.82		
Response *n*(%)			.37	
1	1 (2%)	1 (2%)		
2	2 (4%)	1 (2%)		
3	8 (16%)	8 (14%)		
4	27 (54%)	30 (51%)		
5	12 (24%)	18 (31%)		
Instructor				
n	47	57		
Mean	4.02	4.25		(-0.09, 0.54)
S.D.	0.94	0.66		
Response *n*(%)			.16	
1	1 (2%)	0 (0%)		
2	1 (2%)	0 (0%)		
3	11 (24%)	7 (12%)		
4	17 (36%)	29 (51%)		
5	17 (36%)	21 (37%)		

[a] Mean score test (Stokes, Davis, & Koch, 1995).

[b] Normal approximation (Spring 2005 minus Fall 2004).

4.5. OTHER OBSERVATIONS

Several other observations regarding the distance version of the course are worth noting, including organizational, instructor, and student issues.

Organizational Issues Purchasing the course text and appropriate software for the course was more difficult for distance students as they did not have access to the campus book store. We identified websites (e.g., software company websites) from which students could purchase the necessary materials, and provided links to these websites from the course website.

Students had different computing platforms, as well as different versions of software. Documents posted on the course website contained special characters/symbols (e.g., statistical notation such as Greek letters, etc.). Some students had difficulty with their software correctly recognizing statistical notation due to the different software versions. Creating pdf files alleviated the problem; however, students still had to download appropriate fonts to view the lecture notes.

Instructor Issues There was a substantial increase in the volume of e-mail for the instructor and the teaching assistants. This increase could be due in part to the increase in enrollment, but may also be due to the fact that e-mail is the only method of communication for the distance students or that iterative communication between students and instructors/TAs is necessary for biostatistics courses. Homework submissions via e-mail can also get blocked by spam filters.

In traditional courses, homework is submitted in class by all students, and thus an instructor has all student homework organized into a single location. However, with the distance education option, there was an increase in the proportion of homework assignments that were submitted by fax and e-mail, making it more difficult to keep track of submitted homework. This may be more of an issue in courses such as biostatistics, when frequent homework assignments are part of the course structure and it is necessary to regularly provide feedback to students.

The instructor learned to avoid phrases such as "over here" when using a laser pointer during lectures, as distance students were not able to view where the instructor was pointing. Instead, the instructor used phrases such as "in the upper right-hand corner." This can be potentially problematic with graphs, figures, and other non-text slides often used in biostatistics courses to illustrate concepts and visualize analyses.

Student Issues Homework was returned to distance students using postal mail, and thus could take a couple of weeks to reach distance students that lived outside of the United States. A few distance students noted that this was a significant disadvantage.

Distance students cannot attend help sessions. Although access to the instructor and all of the TAs was available during the course, some distance students believed that not being able to attend help sessions was a significant disadvantage. This can be potentially problematic when students have to learn how to use statistical software.

Some students noted that something was lost when watching the video versus the live lecture, comparing it to watching a concert live versus watching it on television.

A few semi-local students stated that the availability of the video provided time to study that would otherwise be spent commuting. Notably, commuting to live lectures for our course involves commuting and parking in a congested area (Harvard Square in Cambridge, Massachusetts, USA).

5. DISCUSSION

In this study, we observed an increase in enrollment in an "Introduction to Biostatistics" course offered with a distance option. Notably, enrollment doubled from the previous semester, with the "hybrid" option appearing particularly attractive to students. We failed to identify a significant distance effect with regard to student grades. We found a statistically significant video effect with respect to exam grades: Traditional students in a course that offered a distance option performed worse than traditional students in a purely traditional course. However, we believe that these differences in grades are likely due to unintentional semester-to-semester variability in exam difficulty, but cannot rule out a potential detrimental effect of the distance option on traditional students. We note, however, that no differences between overall grade point averages were detected. No evaluation differences were noted between the semesters that offered a distance option compared to the semester that offered the course only traditionally.

We acknowledge possible limitations of this study due to several potential sources of bias. Because this is a non-randomized study, it is possible that comparison groups were different at baseline (selection bias). Our data do not rule out the possibility that either

(1) non-traditional students are superior students but distance delivery is inferior, or (2) non-traditional students are inferior students but distance delivery is superior. The "distance effect" is not solely the distance delivery effect, but is actually a combination of this effect and differences between traditional and non-traditional students. If there are differences between traditional and non-traditional students, then it is difficult to estimate one important parameter of interest: the difference between how a student would perform in the course from a distance versus how the student would perform in the course if they took the course traditionally. In an attempt to identify dissimilarities between traditional and non-traditional students, we collected "baseline" data using the questionnaire and failed to find significant differences between comparison groups for some important variables. However, group differences may exist with respect to other important characteristics that we were not able to identify or measure. Many distance students do not have the option to take the course traditionally due to geographic or time constraints, and for these students, the important question is whether they can successfully complete the course. It is also possible that our reference group (i.e., traditional students) is not representative of all traditional students. We are able to study only the students that happen to enroll into our course. We also note that the semester-to-semester variability in exams and assignments, and the tendency for an instructor to revise his or her teaching methods over time, could affect the between semester comparisons. Furthermore, because we were able to use data only from students that signed the Research Questionnaire (52/71, 73%, in Fall 2004 and 79/140, 56%, in Spring 2005), it is possible that a volunteer bias could affect group comparisons of baseline data. Another limitation is that we do not have data regarding classroom attendance. Lastly, it is important to note the possibility of informative drop-out. For example, if all students that perform poorly early in the course drop out, leaving only the students that perform well (and this drop-out is more prevalent in one comparison group than another), then group comparisons would be biased. However, we have no data to conduct such an evaluation.

We further stress that our results should not be generalized broadly due to the differences in distance methods of delivery. More studies will be needed to investigate whether our results may be generally applicable to "Introduction to Biostatistics" courses with asynchronous video as an option, or if these results apply only to our university. Our study is not definitive, but it is a first step in providing valuable information that can be used as a basis for future research.

Based on our experience, we offer the following recommendations and comments for instructors of distance courses:

1. Instructors may wish to consider themselves students of the distance education process, and attempt to search for ways to understand the unique method of delivery and communication, and how it affects student learning. Attempt to identify ways to help students learn under this system. Learning statistical software can be particularly difficult for students without timely feedback.
2. Provide avenues of communication, such as e-mail, for which relatively timely responses can be delivered. For example, if you have very experienced TAs (as in this course), then assign specific TAs to communicate with distance students via e-mail. This approach should be pursued cautiously, as more instructor control and oversight will likely be required with inexperienced TAs. Whether e-mail is directed by the instructor or TAs, it is important to regularly respond to e-mail. Several web-based course management tools, such as WebCT and Blackboard, can be helpful as bulletin boards. Chat rooms, wikis, and blogs can also allow and facilitate

class discussions. Additionally, on-line testing and students' performance feedback can be done with the use of these tools.

3. Create a comprehensive website that provides timely course materials and answers to common questions using software tools, such as WebCT or Blackboard. Providing handouts for statistical software is helpful for students. This will help to reduce the number of student questions.
4. Make sure that there is adequate school support, including appropriate technology with high quality sound and video, as well as support for distributing homework and for arranging of proctored exams.
5. Be prepared for more work. Allow more time for preparation and interaction with distance students. (It is thus appropriate for the instructors and TAs to be appropriately compensated. The Harvard Extension School increases compensation for distance education courses.)
6. Learn to be flexible and adaptable to distance student needs. Create a balanced approach of understanding the potential difficulties for distance students, and maintaining academic integrity, by requiring students to be responsible with assignments and exams.
7. Discuss the ownership and future use of videotapes with the school. Legality questions may arise regarding the future use of such videotapes.

Distance education remains a controversial topic. However, distance education is here to stay for the foreseeable future. More extensive evaluations of the effectiveness of distance learning through comparisons of distance versus traditional students using internal control groups are needed. Such evaluations may be difficult as the distinction between online and traditional courses begins to blur; many traditional courses are also beginning to incorporate more online components such as message boards, chat rooms, and the electronic filing of homework. Evaluations of other biostatistics/statistics courses, and other methods of delivery, are also needed. Methods to enhance student learning in such courses should be researched. Instructors should realize that they are also students and, therefore, need to learn how to teach in the realm of distance education.

REFERENCES

Allen, I. E.., & Seaman, J. (2005). *Growing by degrees: Online education in the United States, 2005.* Wellesley, MA: The Sloan Consortium.

Allen, M., Bourhis, J., Burrell, N., & Mabry, E. (2002). Comparing student satisfaction with distance education to traditional classrooms in higher education: A meta-analysis. *The American Journal of Distance Education, 16*(2), 83-97.

Bourne, J. R., McMaster, E., Rieger, J., & Campbell, J. O. (1997). Paradigms for on-line learning: A case study in the design and implementation of an asynchronous learning networks (ALN) course. *Journal of Asynchronous Learning Networks, 1*(2).
[Online: www.sloan-c.org/publications/jaln/v1n2/]

Bruce, J. C., Bond, S. T., & Jones, M. E. (2002). Teaching epidemiology and statistics by distance learning. *Statistics in Medicine, 21*, 1009-1020.

Carr, S. (2001). Union publishes guide citing high cost of distance education. *Chronicle of Higher Education, 47*(35), 39-41.

Clow, K. E. (1999). Interactive distance learning: Impact on student course evaluations. *Journal of Marketing Education, 21*(2), 97-105.

Foster, B. (2003). On-line teaching of mathematics and statistics. *Teaching Mathematics and its Applications, 22*(3), 145-153.

Gagne, M., & Shepherd, M. (2001). A comparison between a distance and a traditional graduate accounting class. *T.H.E. Journal, 28*(9), 58-65.

Gilmour, H. (2002). Discussion: Teaching epidemiology and statistics by distance learning. *Statistics in Medicine, 21*, 1021-1022.

Harrington, D. (1999). Teaching statistics: A comparison of traditional classroom and programmed instruction/distance learning approaches. *Journal of Social Work Education, 35,* 343-352.

Hillstock, L. G. (2005). A few common misconceptions about distance learning. In P. Smith & C. Smith (Eds.), *Campus Technology: Anticipating the Future. Proceedings of the Association of Small Computer Users in Education Annual Conference* (pp. 138-145). Myrtle Beach, SC: ASCUE.

Hiltz, S. R., Coppola, N., Rotter, N., & Turoff, M. (2001). Measuring the importance of collaborative learning for the effectiveness of ALN: A multi-measure, multi-method approach. *Journal of Asynchronous Learning Networks, 4*(2). [Online: http://www.sloan-c.org/publications/jaln/v4n2/index.asp]

Hodges, J. L., & Lehmann, E. L. (1963). Estimation of location based on rank test. *The Annals of Mathematical Statistics, 34*, 598-611.

Johnson, S. D., Aragon, S. R., Shaik, N., & Palma-Rivas, N. (1999). Comparative analysis of online vs. face-to-face instruction. In P. De Bra & J. Leggett (Eds.), *Proceedings of the WebNet 99 World Conference on the WWW and Internet* (pp. 581-586). Charlotesville, VA: Association for the Advancement of Computing in Education

Katz, Y. J., & Yablon, Y. B. (2003a). Locus of control, self-esteem, motivation and satisfaction, teaching and learning through an internet-based and traditional 'Introduction to Statistics' course. *The Quality Dialogue — Integrating Quality Cultures in Flexible, Distance and eLearning: Proceedings of European Distance Education Network* (pp. 183-187). Rhodes, Greece: EDEN.

Katz, Y. J., & Yablon, Y. B. (2003b). Online university learning: Cognitive and affective perspectives. *Campus Wide Information Systems, 20*(2) 48-54.

Kendall, M. (2001). Teaching online to campus-based students: The experience of using WebCT for the community information module at Manchester Metropolitan University. *Education for Information, 19*(4), 325-346.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* San Francisco: Holden-Day.

Li, H. (2002, March). *Distance education: Pros, cons, and the future*. Paper presented at the Annual Meeting of the Western States Communication Association, Instructional Division, Long Beach, CA.

McGready, J. (2006, August). *Basic biostats: Online learning versus onsite learning*. Paper presented at the Joint Statistical Meetings, Seattle, WA.

Merisotis, J. P., & Phipps, R. A. (1999). What's the difference? Outcomes of distance vs. traditional classroom-based learning. *Change, 31*(3), 12-17.

Ng, K. (2000). Costs and effectiveness of online courses in distance education. *Open Learning, 15*(3), 301-308.

Pagano, M., & Gauvreau, K. (2000). *Principles of biostatistics* (2nd ed.). Pacific Grove, CA: Duxbury.

Phelps, R. H., Wells, R. A., Ashworth, R. L., & Hahn, H. A. (1991). Effectiveness and costs of distance education using computer-mediated communication. *American Journal of Distance Education, 5*(3), 7-19.

Phipps, R., & Merisotis, J. (1999). *What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education.* Washington, DC: The Institute of Higher Education Policy.
[Online: http://www.ihep.com/Pubs/PDF/Difference.pdf]

Rivera, J., McAlister, M., Khris, R., & Margaret, L. (2002). A comparison of student outcomes and satisfaction between traditional and web based course offerings. *Online Journal of Distance Learning Administration, 5*(3).
[Online: www.westga.edu/~distance/ojdla/fall53/rivera53.html]

Rooney, P., Hussar, W., Planty, M., Choy, S., Hampden-Thompson, G., Provasnik, S., & Fox, M. A. (2006). *The Condition of Education, 2006 (NCES 2006-071),* U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Russell, T. L. (1999). *The no significant difference phenomenon.* Raleigh, NC: North Carolina State University.

Speed, F. M., & Hardin, H. (2001). Teaching statistics via distance: Duplicating the classroom experience. *Communications in Statistics: Simulation and Computation, 30*(2), 391-402.

Stephenson, R. W. (2001). Statistics at a distance. *Journal of Statistics Education, 9*(3).
[Online: www.amstat.org/publications/jse/v9n3/stephenson.html]

Stokes, M. E., Davis, C. S., & Koch, G. G. (1995). *Categorical data analysis using the SAS system.* Cary, NC: SAS Institute, Inc.

Valentine, D. (2002). Distance learning: Promises, problems, and possibilities. *Online Journal of Distance Learning Administration, 5*(3).
[Online: www.westga.edu/~distance/ojdla/fall53/valentine53.html]

Wegman, E. J., & Solka, J. L. (1999). Implications of distance learning methodologies for statistical education. *1999 Proceedings of the American Statistical Association,* Sections on Statistsical Education, Teaching Statistics in the Health Sciences, and Consulting (pp. 13-16). Alexandria, VA: American Statistical Association.

Wernet, S. P., Olliges, R. H., & Delicath, T. A. (2000). Postcourse evaluations of WebCT (Web Course Tools) classes by social work students. *Research on Social Work Practice, 10*(4) 487-504.

SCOTT R. EVANS
FXB 513
Department for Biostatistics
Harvard School of Public Health
651 Huntington Ave
Boston, MA 02115

A STRUCTURAL EQUATION MODEL ANALYZING THE RELATIONSHIP OF STUDENTS' ATTITUDES TOWARD STATISTICS, PRIOR REASONING ABILITIES AND COURSE PERFORMANCE

DIRK T. TEMPELAAR
Maastricht University, The Netherlands
D.Tempelaar@ke.unimaas.nl

SYBRAND SCHIM VAN DER LOEFF
Maastricht University, The Netherlands
S.Loeff@ke.unimaas.nl

WIM H. GIJSELAERS
Maastricht University, The Netherlands
W.Gijselaers@erd.unimaas.nl

ABSTRACT

Recent research in statistical reasoning has focused on the developmental process in students when learning statistical reasoning skills. This study investigates statistical reasoning from the perspective of individual differences. As manifestation of heterogeneity, students' prior attitudes toward statistics, measured by the extended Survey of Attitudes Toward Statistics (SATS), are used (Schau, Stevens, Dauphinee & DeVecchio, 1995). Students' statistical reasoning abilities are identified by the Statistical Reasoning Assessment (SRA) instrument (Garfield 1996, 1998a, 2003). The aim of the study is to investigate the relationship between attitudes and reasoning abilities by estimating a full structural equation model. Instructional implications of the model for the teaching of statistical reasoning are discussed.

Keywords: *Statistics education research; Statistical reasoning; Achievement motivations; SATS; SRA; Structural equation modelling*

1. INTRODUCTION

Recent research into statistical reasoning about variation, distribution, and sampling distributions has created important insights into the developmental process of statistical reasoning skills. Most research has focused on the identification of subsequent, hierarchically-ordered stages of reasoning development by means of qualitative research methods such as thinking-aloud sessions and in-depth interviews. Two recent special issues of this journal (*SERJ*, Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005) and an edited volume (Ben-Zvi & Garfield, 2004a) contain a wealth of such empirical studies into the cognitive process of developing reasoning abilities and of instructional tools that might foster these developments. The present research investigates statistical reasoning from a somewhat different perspective. It examines individual differences among students learning statistics and statistical reasoning. These individual differences demonstrate much variability: Students enter learning processes with different background characteristics and different perceptions of the learning context. As a manifestation of students' heterogeneity, this study uses students' prior attitudes toward statistics. The main aim of this study is to investigate the relationship between students' attitudes toward statistics and their prior statistical reasoning abilities when entering an introductory statistics course.

Contemporary research in statistics education distinguishes an array of different but related cognitive processes in learning statistics: statistical literacy, statistical reasoning, and statistical thinking. See for example the special section of the *Journal of Statistics Education* (Short, 2002), the two special issues of *SERJ* (Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005), Ben-Zvi and Garfield (2004a), and Pfannkuch and Wild (2004). The demarcation of these three cognitive processes not being complete, it is well accepted that statistical literacy represents the most basic skills (Ben-Zvi & Garfield, 2004c). Gal (2004) distinguishes two interrelated components in statistical literacy: the ability to "interpret and critically evaluate statistical information, data-related arguments, and stochastic phenomena," and the ability to "discuss or communicate" these (see also Rumsey, 2002). Statistical reasoning is the ability to "explain why a particular result is expected or has occurred, or explain why it is appropriate to select a particular model or representation" (delMas, 2004a; see also Garfield & Chance, 2000; Garfield, 2002). Statistical thinking involves an "understanding of why and how statistical investigations are conducted and the 'big ideas' that underlie statistical investigations" (Ben-Zvi & Garfield, 2004; see also Pfannkuch & Wild, 2004; Chance, 2002). Literacy, reasoning, and thinking are to some extent achieved even before formal schooling in statistics takes place. Those naïve conceptions learned outside school can be correct or incorrect in nature.

In the 1970s, cognitive research into statistical and probabilistic reasoning revealed several categories of fallacies in human reasoning, with examples such as the 'Law of small numbers,' the 'Representativeness misconception' (Kahneman, Slovic, & Tversky, 1982), the 'Outcome orientation' (Konold, 1989), and the 'Equiprobability bias' (Lecoutre, 1992). Most of that research is documented in the seminal work of Kahneman et al. (1982), as cited in Garfield and Ahlgren (1988). In the decades thereafter, following the reform movement in statistics education, research shifted its focus from probabilistic reasoning to reasoning with data (Pfannkuch & Wild, 2004), as evidenced in the topics of the recent series of SRTL research forums and the compilation of their major contributions in Ben-Zvi and Garfield (2004a).

Another important development in recent decades is the design of assessment instruments for statistical literacy, reasoning, and thinking (delMas, 2002; Garfield & Ben-Zvi, 2004a). Paraphrasing Chance (2002), 'if not assessed, it cannot be valuable,' and assessment instruments were needed to match the focus on literacy, reasoning, and thinking. Several instruments also grew out of the need for assessment tasks that could be used in the context of research projects. Quantitative assessment instruments are still scarce, and are all derived from the first and most prominent instrument in the field: Statistical Reasoning Assessment (SRA). The SRA was developed by Konold and Garfield (Konold, 1989; Garfield, 1996, 1998a, 2003) as part of a project evaluating the effectiveness of a new statistics curriculum in U.S. high schools. The instrument is based on the well-described classes of misconceptions and their antipodes, the learned or unlearned correct conceptions, that emerged from the cognitive science research into reasoning fallacies (Garfield, 2003; Garfield & Ahlgren, 1988). In current terminology – the SRA was developed long before recent discussions on the demarcation of literacy, reasoning, and thinking – fallacies addressed in the SRA are of all three types. Being designed in the earlier stages of the reform movement in statistics education (Ben-Zvi & Garfield, 2004c), the SRA focuses both on statistical and probabilistic reasoning. Newer assessment instruments, related to the SRA but focusing more strongly on reasoning with data, are currently being developed in the framework of the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project (delMas, 2004b; see also https://app.gen.umn.edu/artist/). As newer instruments were not yet available, the SRA was the most appropriate tool at the time of this study to assess students' reasoning abilities in the large-scale applications typical of educational practice.

Empirical studies on statistical reasoning focus predominantly on the cognitive developmental process students go through when learning reasoning abilities, and on the instructional tools that may foster these developments. The large majority of these studies are empirical in nature in that they use descriptions, often achieved by thinking-aloud sessions or interviews of the cognitive states of students, to reconstruct a developmental trajectory (Ben-Zvi & Garfield, 2004a). Garfield and Ben-Zvi (2004b, p. 399) ascertain "It may seem strange, given the quantitative nature of statistics, that most of the studies … include analyses of qualitative data, particularly videotaped observations or interviews." Yet such studies allow identification of different states of students' reasoning abilities and subsequent stages in the developmental process. Our study chooses a different perspective based on individual differences in student-related factors by investigating the role of non-cognitive individual differences in the cognitive development of students. This type of study has, at least in the context of statistics and mathematics education, a long tradition (Gal & Garfield, 1997; McLeod, 1992). In conceptualizing the non-cognitive domains of education, McLeod (1992) distinguishes among emotions, attitudes and beliefs. In most studies of learning processes in statistical education, the focus is on beliefs and attitudes, rather than emotions; see for example Gal and Ginsburg (1994) and Gal and Garfield (1997). Probably the best known, and certainly most validated, model on the role of attitudes in learning statistics is the model developed by Schau and co-authors (Schau, Stevens, Dauphinee & DeVecchio 1995). The Schau-model is based on the expectancy-value model for achievement motivations designed by Eccles and Wigfield (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000, 2002). In that model, students' expectancies for success and the value they contribute to succeeding are important determinants of their motivation to perform achievement tasks. Expectancy for success crystallizes in two different concepts: belief in one's own ability to perform a task, and a perception of the task demand. Subjective task value is generally modeled in a single concept, comprising several aspects: attainment value (importance of doing well on a task), intrinsic value (interest in and enjoyment gained from doing the task), utility value (usefulness), and costs (spent efforts) (Eccles, 2005). The contribution of Schau and co-authors to the development of the expectancy-value model of achievement motivations is two-fold. First, they designed the SATS measurement instrument to adapt the generic expectancy-value model to the statistical domain (Schau et al., 1995; Dauphinee, Schau & Stevens, 1997). Second, they extended the generic model by introducing new concepts obtained by disentangling the broad task-value concept of the expectancy-value model. In the first 28-item version of SATS, the task-value concept is broken up into an affective concept, focusing most on the enjoyment aspect of intrinsic values, and a valuation concept, focusing on the remaining components of attainment and utility values. The model of the first version thus contains two expectancy factors that deal with students' beliefs about their own ability and perceived task difficulty, Cognitive Competence and Difficulty, and two subjective task-value concepts that encompass students' feelings toward and attitudes about the value of the subject, Affect and Value (Schau, 2003). Empirical research, both within the statistics domain (Dauphinee et al., 1997; Sorge & Schau, 2002; Hilton, Schau, & Olsen, 2004) and in other academic domains (Tempelaar, Gijselaers, Schim van der Loeff, & Nijhuis, 2007) supports the distinction of these affective and valuation aspects. In a second, 36-item version of SATS (C. Schau, personal communication, November 30, 2003), two more concepts are introduced: Interest and Effort. The Interest concept shapes the interest aspect of the intrinsic value component in the expectancy-value model, whereas the Effort concept shapes the perceived costs component in the subjective task-value (Eccles, 2005). To the knowledge of the authors, no empirical studies based on the extended SATS instrument have yet been published. Empirical studies of the 28-item version of SATS, referred to above, focus on the structure of attitudes alone, or on the structure of attitudes in relation to statistics course performances. The context of these studies is thereby slightly different from most studies in the expectancy-value framework that focus primarily on the relation between attitudes and learning task choices (such as course selection) rather than learning task outcomes.

The main contribution of this paper is to investigate the dependency of students' prior reasoning abilities on their attitudes toward statistics when entering an introductory statistics course. In the formulation of this research question, attitudes are hypothesized to be causal to statistical reasoning abilities. The hypothesized direction of causality is in agreement with process models of learning (see for example Garfield, Hogg, Schau, & Whittinghill, 2002), in which affective, student-related factors are regarded as determinants for cognitive, learning-outcome-related factors. In addition, attitudinal variables possess a trait-like nature, in contrast to reasoning abilities that possess a state-like nature. Therefore, the hypothesized causal direction follows the general modeling pattern of stable traits determining malleable states. In order to do so we start the empirical third section by developing confirmatory latent factor models for attitudes, based on the extended SATS instrument, and for statistical reasoning, based on the SRA instrument. Subsequently, these factor models are integrated into a full structural equation model that explains reasoning abilities by attitude factors. To be able to put this relationship into perspective, two further cognitive constructs are added to this model: course performance measured by quiz and final exam scores. This extension allows characterizing reasoning abilities not only by their direct relationship with attitudinal variables, but also by a comparison of that relationship with the ones between attitudes and course performances.

One of the implications of our model is that where different learning approaches provide alternative routes to achieve traditional course performances, perhaps one more efficiently than the others but all contributing to the same learning goal, this seems not to be true for statistical reasoning abilities. Some learning approaches really hinder achievement of reasoning skills. The model outcomes thus have strong implications for the development of instructional programs in statistical reasoning, which is one of the topics discussed in the concluding section.

2. METHOD

2.1. PARTICIPANTS AND PROCEDURE

In this study, the statistical reasoning of students participating in the "International Business" and "International Economics" programs of the Maastricht University was investigated. A large number of students, 842 and 776 respectively, from these two programs participated in the first year, first semester course Quantitative Methods (QM) in 2004/05 and 2005/06. This is a compulsory introduction to mathematics and statistics for all students. Of these 1618 students, 64% were male and 36% were female. Another relevant decomposition was that 39% students had a Dutch secondary school diploma, versus 61% students with non-Dutch diplomas (most of them of German nationality).

Part of the data analyzed in this study comes from regular student quizzes and examinations. In the QM course, three assessment instruments are applied. One is a final exam, in multiple-choice format, covering both statistics and mathematics. Items in the exam focus on students' ability to apply statistical and mathematical methods; those in statistics are motivated by the Advanced Placement Statistics exams (e.g., http://apcentral.collegeboard.com). Secondly, both for statistics and mathematics, three quizzes are taken spread over the eight weeks of the course. Quizzes are optional; they give rise to bonus points for the exam score. In practice, all students participate in most of the quizzes. For this study, quiz scores are aggregated over the three quizzes. The third assessment instrument is a student project. For this project, students collect personal data by completing several self-report instruments concerning their study approach and preferred strategies. Later on, they perform an explorative analysis of these data. Students are informed that the self-reported data are also used for three additional purposes: to provide study advice to students who have adopted an inefficient study approach, for course-improvement purposes, and for research. The project is compulsory, and assessed with pass/fail. Because students can acquire feedback on their project in several stages of its development, the final assessment of it is not very informative, and is not included in this study.

The SATS and the SRA were the first self-report instruments to be administered during the first days of the course. Responses to both surveys therefore reflect students' prior attitudes and beliefs toward statistics and their prior reasoning abilities. Scores cannot be influenced by impressions of the educational process, nor by knowledge achieved in the course itself.

Both instruments are quantitative in nature, and generate observations that can be regarded as proxies for the underlying, but unobservable, theoretical constructs. Therefore, the investigation of the relationship between attitudes and reasoning abilities requires the estimation of two confirmatory latent factor models for attitudes on the one side, and for statistical reasoning on the other, as well as the integration of both these factor models into a full structural equation model. To this model, we add two indicators of course performance: latent variables measuring the strongly cognitive-based scores in the final exam, and the more effort-based scores in quizzes. The primary reason for doing so is that it allows for characterization of the particular position statistical reasoning takes within the spectrum of different performance indicators.

2.2. MEASURES

Statistical reasoning abilities The Statistical Reasoning Assessment (SRA) is a test consisting of 20 multiple-choice or multiple-answer items developed by Konold and Garfield as part of a project evaluating the effectiveness of a new statistics curriculum in U.S. high schools (Konold, 1989; Garfield, 1996, 1998a, 2003). Each item in the SRA describes a statistics or probability problem and offers four to eight choices of responses. Most responses include a statement of reasoning, explaining the rationale for a particular choice. For every item, one response corresponds to a category of correct reasoning; all or most of the other responses correspond to categories of misconceptions. For a full description of the individual items and the eight correct reasoning scales and eight misconceptions scales, see Garfield (1998a, 2003); Table 1 summarizes the scales of the description of the individual items and the eight correct reasoning scales and eight

Table 1. SRA Correct reasoning scales and misconceptions scales; based on Garfield (2003).

Correct Reasoning Scales:

Prob: Correctly interprets probabilities. Assesses the understanding and use of ideas of randomness and chance to make judgments about uncertain events.

Aver: Understands how to select an appropriate average. Assesses the understanding of what measures of center tell about a data set, and which are best to use under different conditions.

Comp: Correctly computes probability, both understanding probabilities as ratios, and using combinatorial reasoning. Assesses the knowledge that in uncertain events not all outcomes are equally likely, and how to determine the likelihood of different events using an appropriate method.

Indep: Understands independence.

Sampl: Understands sampling variability.

Correl: Distinguishes between correlation and causation. Assesses the knowledge that a strong correlation between two variables does not mean that one causes the other.

2Way: Correctly interprets two-way tables. Assesses the knowledge of how to judge and interpret a relationship between two variables, knowing how to examine and interpret a two-way table.

LrgS: Understands the importance of large samples. Assesses the knowledge of how samples are related to a population and what may be inferred from a sample; knowing that a larger, well-chosen sample will more accurately represent a population; being cautious when making inferences made on small samples.

Misconception scales:

AverMc: Misconceptions involving averages. This category includes the following pitfalls: believing averages are the most common number; failing to take outliers into consideration when computing the mean; comparing groups based on their averages only; and confusing mean with median.

OutcO: Outcome orientation. Students use an intuitive model of probability that leads them to make yes or no decisions about single events rather than looking at the series of events; see Konold (1989).

High%: Good samples have to represent a high percentage of the population. Size of the sample and how it is chosen are not important, but it must represent a large part of the population to be a good sample.

Small: Law of small numbers. Small samples best resemble the populations from which they are sampled, so are to be preferred over larger samples.

Repre: Representativeness misconception. In this misconception the likelihood of a sample is estimated based on how closely it resembles the population. Documented in Kahneman, Slovic, & Tversky (1982).

Cause: Correlation implies causation.

EquiPr: Equiprobability bias. Events of unequal chance tend to be viewed as equally likely; see Lecoutre (1992).

Groups: Groups can be compared only if they have the same size.

description of the individual items and the eight correct reasoning scales and eight misconceptions scales, see Garfield (1998a, 2003); Table 1 summarizes the scales of the instrument. In the design process of the instrument, the authors included several stages directed at achieving good validity and reliability. With regard to criterion-related validity, Garfield (2003) reports extremely low correlations with different course outcomes, suggesting statistical reasoning and misconceptions are unrelated to course performance. In addition, Garfield (2003) reports satisfactory test-retest reliabilities, but low internal consistency reliability coefficients, implying that scales and misconception scales respectively appear not to measure one single ability or trait.

In terms of the classification into the more recently developed categories of statistical literacy, reasoning, and thinking, the allocation of individual reasoning abilities and misconceptions to these three classes is not obvious. Aver, TWay, AverMc, High%, and Groups refer to basic data-related skills, and seem to fit best in the literacy category. At the other extreme, Comp, Sampl, Correl, Small, Cause, and EquiPr involve probability and statistical theory related concepts, and might better suit the thinking category. The remaining scales, referring to notions of probability and uncertainty, would then fit the reasoning category. We return to this issue when discussing descriptive statistics of SRA data obtained from this study and a limited number of other studies that provide empirical data on the instrument: Garfield (1998b, 2003), Garfield and Chance (2000), Liu (1998) and Sundre (2003).

Attitudes and beliefs toward statistics Attitudes are measured with the Survey of Attitudes Toward Statistics (SATS) developed by Schau and co-authors (Schau et al., 1995; Dauphinee et al., 1997). There are two existing versions of the SATS, both consisting of seven-point Likert-type items measuring aspects of post-secondary students' statistics attitudes. The 28-item version of SATS contains four scales, as indicated below. Each scale is accompanied by two examples of items, one positively and one negatively worded:

- Affect (six items) - measuring positive and negative feeling concerning statistics, the enjoyment aspect of intrinsic value: *I like statistics*; *I am scared by statistics*.
- Cognitive Competence (six items) - measuring attitudes about intellectual knowledge and skills when applied to statistics, the self-concept of one's ability component in the expectancy-value model: *I can learn statistics*; *I have no idea of what's going on in statistics*.
- Value (nine items) - measuring attitudes about the usefulness, relevance, and worth of statistics in personal and professional life, the utility and attainment components of task value: *I use statistics in my everyday life*; *I will have no application for statistics in my profession*.
- Difficulty (seven items) - measuring attitudes about the difficulty of statistics as a subject, the perception of the task demand: *Statistics formulas are easy to understand*; *Statistics is highly technical*.

Schau et al. (1995), Dauphinee et al. (1997), and Harris and Schau (1999) elaborate on the development process of the instrument. The instrument is freely available from the internet (Schau, Dauphinee, Del Vecchio, & Stevens, 1999). Validation research in two very large samples of undergraduate students has shown that a four-factor structure provides a good description of responses to the SATS-instrument (Dauphinee et al., Hilton et al., 2004).

Recently, Schau has developed a 36-item version of the SATS, containing two additional scales, each covered by four, positively worded, items (Schau, personal communication, November 30, 2003). These scales, with one item example, are

- Interest (four items) - students' level of individual interest in statistics, the interest aspect of intrinsic value: *I am interested in learning statistics*.
- Effort (four items) - amount of work the student expends to learn statistics, the perceived cost component of task value: *I plan to work hard in my statistics course*.

2.3. DATA ANALYSIS

Parceling The very first step in the data analysis is to reverse the negatively worded items in the SATS instrument, such that for all items a higher score corresponds to a more positive attitude. This step is worthwhile to mention because it requires attentiveness in the interpretation of the construct Difficulty. High scores for Difficulty express a more positive attitude, implying that a better name for the Difficulty scale would have been 'perceived lack of difficulty.' The second step of analysis is the parceling of the SATS data, following earlier empirical work by Schau and co-authors (Schau et al., 1995; Dauphinee et al., 1997; Hilton et al., 2004). The technique of item parceling, where items from the same subscale are aggregated into several parcels or miniscales, has been adopted in empirical studies for several reasons: to obtain more continuous and normally distributed observed data, to reduce the number of model parameters to achieve a more attractive variable to sample size ratio, and to get more stable parameter estimates (Bandalos, 2002; Hau & Marsh, 2004; Marsh, Hau, Balla, & Grayson, 1998).

In parceling items, Hau and Marsh (2004) advise not to reduce the number of indicators for each latent construct beyond a minimum of three. Next, they recommend to counterbalance skewness in the presence of strong non-normality by creating parcels out of item pairs with opposite skew. In order to determine the relevance of this recommendation of counterbalancing skewness for our data set, the degree of non-normality of the data was calculated as a preliminary step to parceling. In the data of the first four SATS factors, no indications of non-normality were found in any of the self-reported questionnaires beyond Hau and Marsh's (2004) category of 'moderately non-normal,' implying skew = 1.0 and kurtosis = 1.5. Items corresponding to the constructs Interest and especially Effort were however much more strongly skewed.

In the empirical analyses of their 28-item SATS data, Schau et al. (1995), Dauphinee et al. (1997), and Hilton et al. (2004) adopt an item parceling scheme based on balancing with respect to the positively and negatively worded items, size of parcel means, standard deviations, and skew (see Schau et al.). Their parceling solution contains two parcels for Affect, Cognitive Competence, and Difficulty each; only Value contains three. Given the rule of thumb of at least three parcels per factor and the advice to counterbalance skew as much as possible, it was decided to apply a parceling scheme different from Schau and co-authors, based only on skewness, and resulting in exactly three parcels per factor.

Statistical analyses This study integrates several techniques of structural equation modeling (SEM). A SEM model is distinct from a path or regression model in that it hypothesizes that crucial variables, such as attitudes in this study, are not directly observable and are better modeled as latent variables than as observable ones. In doing so, a SEM model makes it possible to distinguish two different types of errors: errors in equations, as does the path model, and errors in the observation of variables. Making this distinction is especially worthwhile when errors in important constructs have rather different sizes. Studying reliabilities of several achievement motivations, and their variation over subjects, suggests that this argument applies to this study. In this study, SEM models were estimated with LISREL (version 8.54) using maximum likelihood estimation. For further discussion of SEM see for example Byrne (1998), Kline (2005), and Schumacker and Lomax (2004).

The standard approach to estimate a SEM distinguishes two steps (Schumacker & Lomax, 2004). In the first phase of the two-step model building approach, measurement models for all latent variables in the model are estimated. Measurement models are in general factor models that allow factors, also called traits, and the uniqueness, that is the errors in indicators, to be correlated. In our study, we need to estimate three of such ‘correlated trait’ (CT), ‘correlated uniqueness’ (CU), and ‘confirmatory factor analysis’ (CFA) models: for the SATS data, for the SRA data, and for course performance data. In the second model building step, the structural part of the SEM is estimated. This structural part specifies the relationships between the independent and dependent latent variables. In contrast to the estimation of the measurement models, the estimation of structural relationships is to some extent explorative in nature. The structural part of the full structural equation model is not a priori restricted, except for several hypotheses with regard to the direction of the relationship. For the estimation of these structural parts, two different model modification procedures are applied. The first is called model trimming (Kline, 2005) or backward search (Schumacker & Lomax, 2004). Starting from a full matrix of structural path coefficients, one by one, parameters are restricted to zero if they prove non-significant, until all remaining structural parameters are significant. The second approach is called model building (Kline, 2005) or forward search (Schumacker & Lomax, 2004). It starts from a zero matrix of structural paths coefficients, and frees parameters one by one, in the order indicated by the value of the modification indices, up to point where no more significant improvement in fit is achieved. Because in both approaches subsequent models are nested, the chi-square difference statistic can be used to assess model fit. In all five subjects, both forward and backward searches converge to the same final model. Model modification is a form of explorative analysis, and brings along the risk of capitalization on chance.

With large sample sizes as in our study, the χ^2 test statistic is known to always reject in any formal test of significance (Byrne, 1998; Marsh & Yeung, 1996). For that reason, and following Marsh and Yeung (1996), and Hilton et al. (2004), emphasis is placed on the Root Mean Square Error of Approximation (RMSEA), the Goodness-of-Fit Index (GFI), the Non-Normed Fit Index (NNFI; termed Tucker-Lewis Index or TLI in Marsh & Yeung, 1996), the Comparative Fit Index (CFI) and the Relative Fit Index (RFI, termed Relative Noncentrality Index or RNI in Marsh & Yeung, 1996), and the normed version of the χ^2 test statistic: χ^2/df. For the last index, no clear-cut guidelines exist; values in the range of 2.0 to 5.0 are acceptable, with lower values indicating better fit. For RMSEA, values ≤ 0.05 indicate good fit, values ≤ 0.08 indicate reasonable fit. The indices GFI, NNFI, CFI, and RFI, all normally lie in the range 0.0 – 1.0, with higher values indicating better fit. As a benchmark for good fit, the value 0.90 is often used (Kline, 2005).

The covariance matrixes required for estimation are available from the authors upon request.

3. RESULTS

3.1. DESCRIPTIVE STATISTICS OF ATTITUDES AND BELIEVES TOWARD STATISTICS

Descriptive statistics of the SATS scales are exhibited in Table 2 and Figure 1. All attitudes are measured using a Likert 1-7 scale. Because all scale means, except for Difficulty, are larger than the neutral value of four, students in our sample express positive attitudes toward statistics for Affect, Cognitive Competence, Value, Interest, and Effort. Means and standard deviations are in line with values reported in Schau (2003) found as pre-test scores in a large class of undergraduate U.S. students; Affect, Cognitive Competence, and Value are slightly more positive in our sample, Difficulty is equal. In comparing our European data with data from U.S. studies, it is important to realize that participants in our study are all in economics and business programs. These programs require students to take math classes in high school through at least intermediate level. Cronbach α reliability coefficients of these four scales are satisfactory, and again in line with intervals of values reported in Schau (2003) from several empirical studies by Schau and co-researchers. No empirical studies exist at this moment that incorporate the new scales of the 36-item SATS version: Interest and Effort. In our study, both these attitudes are clearly positive on average, with (planned) Effort taking a very strong position with a mean of 6.37 on a 1-7 scale. Figure 1 indicates that due to the high scores on Effort, skewness is an issue for this scale, and not for the other scales.

Table 2. Scale means, standard deviations, and Cronbach α's for attitudes toward statistics in our study (n=1458) and as reference, values reported in Schau (2003)

	Mean (Standard deviation)		Cronbach α	
	this study	Schau (2003)	this study	Schau (2003)
Affect	4.52 (1.10)	4.03 (1.14)	0.82	0.80 – 0.89
Cognitive Competence	5.08 (0.89)	4.91 (1.09)	0.78	0.77 – 0.88
Value	5.05 (0.83)	4.86 (1.01)	0.78	0.74 – 0.90
Difficulty	3.59 (0.77)	3.62 (0.78)	0.68	0.64 – 0.81
Interest	5.07 (0.99)		0.80	
Effort	6.37 (0.72)		0.76	

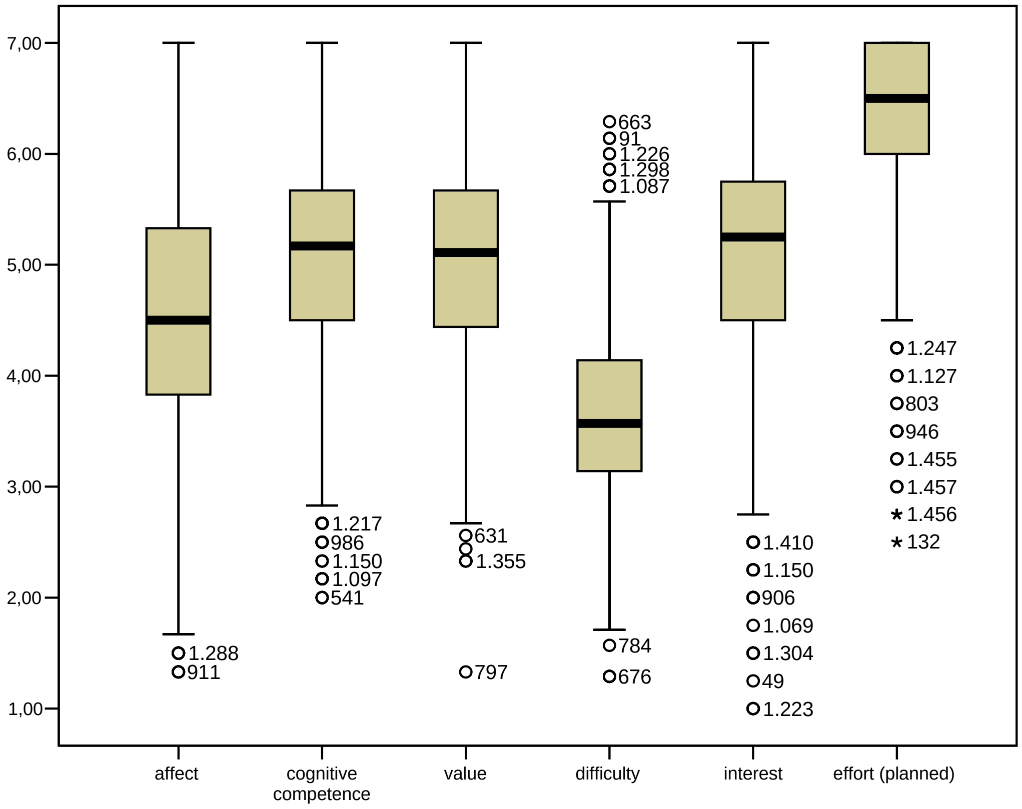


Figure 1. Descriptives of SATS scales (n=1458)

3.2. DESCRIPTIVE STATISTICS OF STATISTICAL REASONING ABILITIES

Descriptive statistics of the SRA data, similar to those reported in Garfield (1998b, 2003), Garfield and Chance (2000) and Liu (1998), are exhibited in Table 3. Because the maximum score of the several scales varies with the total number of answer options corresponding to the scale, the table presents the means of the several scales expressed as a proportion, that is, on a [0-1] scale. In addition to scores on eight reasoning skills, and eight misconceptions, the aggregated correct reasoning score (Correct) and aggregated misconceptions (Misconcep) are reported. The aggregated scores are obtained in the same way as in the studies by Garfield and co-authors by taking the sum over all correct reasoning and misconception items, and re-expressing them as a proportion. Because the number of items per scale ranges from 1 to 5, different scales have a different weight in the total score, so aggregated scores are to be regarded as weighted averages. Data reported by Garfield and co-authors are restricted to means.

Table 3. Scale means and standard deviations for statistical reasoning abilities in our study (n=1499) and as reference, post-course values US college students reported in Garfield (2003)

	Mean (Standard deviation)			Mean (Standard deviation)	
	this study	Garfield (2003)		this study	Garfield (2003)
Prob	0.75 (0.29)	0.68	AverMc	0.46 (0.27)	0.30
Aver	0.71 (0.27)	0.61	OutcO	0.22 (0.17)	0.23
Comp	0.40 (0.25)	0.46	High%	0.15 (0.23)	0.09
Indep	0.64 (0.29)	0.63	Small	0.28 (0.27)	0.29
Sampl	0.28 (0.30)	0.22	Repre	0.12 (0.22)	0.17
Correl	0.66 (0.47)	0.51	Cause	0.28 (0.37)	0.10
Twow	0.74 (0.40)	0.65	EquiPr	0.57 (0.33)	0.56
LrgS	0.71 (0.33)	0.68	Groups	0.29 (0.46)	0.60
Correct	0.58 (0.13)	0.55	Misconcep	0.29 (0.10)	0.27

Outcomes of our study and those reported in Garfield (2003) are remarkably similar, although the composition of groups of participating students is rather different. Garfield’s study refers to U.S. college students surveyed at the end of an introductory course statistics, our study to European university students at the start of such an introductory course. Of the correct reasoning scales, Prob and Twow are amongst those with highest mastery level, and Comp and Sampl with lowest. Of the misconception scales, EquiPr and Groups are high in all studies (in our sample, Groups somewhat less), and High%, Repre and Cause are low.

Conceptions for which we find higher scores than reported in the Garfield studies are Aver, Correl, and Twow. The misconception for which our data indicate a remarkably low relative score is Groups. Of these four scales, three are characterized earlier as being part of the category of statistical literacy. This agrees with the difference in timing of the instrument, as a pre-test in our study, and a post-test in other studies. Not (recently) educated in introductory statistics, it is not surprising that students in our study score relatively high on statistical literacy components, but low on a statistical thinking related component as MC6 (correlation implies causation), typically an important concept to be taught in an introductory course.

As a last observation on average levels of reasoning skills and misconceptions, the high rate of correct answers is noticeable. Of the eight correct reasoning skills, five have means of above 65% correct. Of the eight misconception scales, only two have means larger than 30%.

3.3. MEASUREMENT MODEL OF ATTITUDES AND BELIEFS TOWARD STATISTICS

As a first step in the modeling of the SATS data, an explorative factor analysis was performed (principal components, varimax rotation). The eigenvalue criterion identifies six factors. The scree-criterion demonstrates a large jump at four factors, and a smaller jump at six factors. The newly created scales Interest and Effort clearly qualify as independent factors. The same is true for the scale Value. However, items in the scales Affect, Cognitive Competence, and Difficulty are strongly correlated. This finding coincides with other empirical studies on SATS: Schau et al. (1995), Dauphinee et al. (1997), Hilton et al. (2004), and Cashin and Elmore (2005). On the basis of these high correlations, Cashin and Elmore (2005) decide to reduce the three scales Affect, Cognitive Competence, and Difficulty into one latent factor, whereas in the other three studies they are modeled as separate, but correlated, latent factors. We followed the last approach estimating a six-factor confirmatory factor model on parcelled attitudes data allowing a correlated traits (CT) structure but without cross-loadings in the factor loading matrix and no correlated uniqueness (CU) factor. Table 4 contains fit indices of this CT factor model, Figure 2 the structure of the factor model, including estimated trait correlations.

Table 4. Fit indices of six-factor correlated traits confirmatory factor models of attitudes toward statistics

	χ^2	*df*	RMSEA	GFI	NNFI	CFI	RFI
CT 6CFA model	701.80	123	.057	.95	.97	.97	.96

Fit indices indicate that the hypothesized correlated traits factor model fits the data quite well. Having confirmed the six-factor model, the correlation structure of latent factors depicted in Table 5 deserves prime interest. Table 5 demonstrates that twelve out of fifteen trait correlations are significant. Only three trait correlations appear to be non-significant and are restricted to zero in the estimation of the final version of the factor model, with the other correlations freed.

Table 5. Estimated latent factor correlations of attitudes toward statistics

	Affect	Cognitive Competence	Value	Difficulty	Interest	Effort
Affect	1.00					
Cognitive Competence	0.80	1.00				
Value	0.40	0.43	1.00			
Difficulty	0.61	0.62	-	1.00		
Interest	0.42	0.35	0.63	-	1.00	
Effort	-	0.17	0.34	-0.28	0.44	1.00

Note. All reported correlations are significant at p < 0.000001.

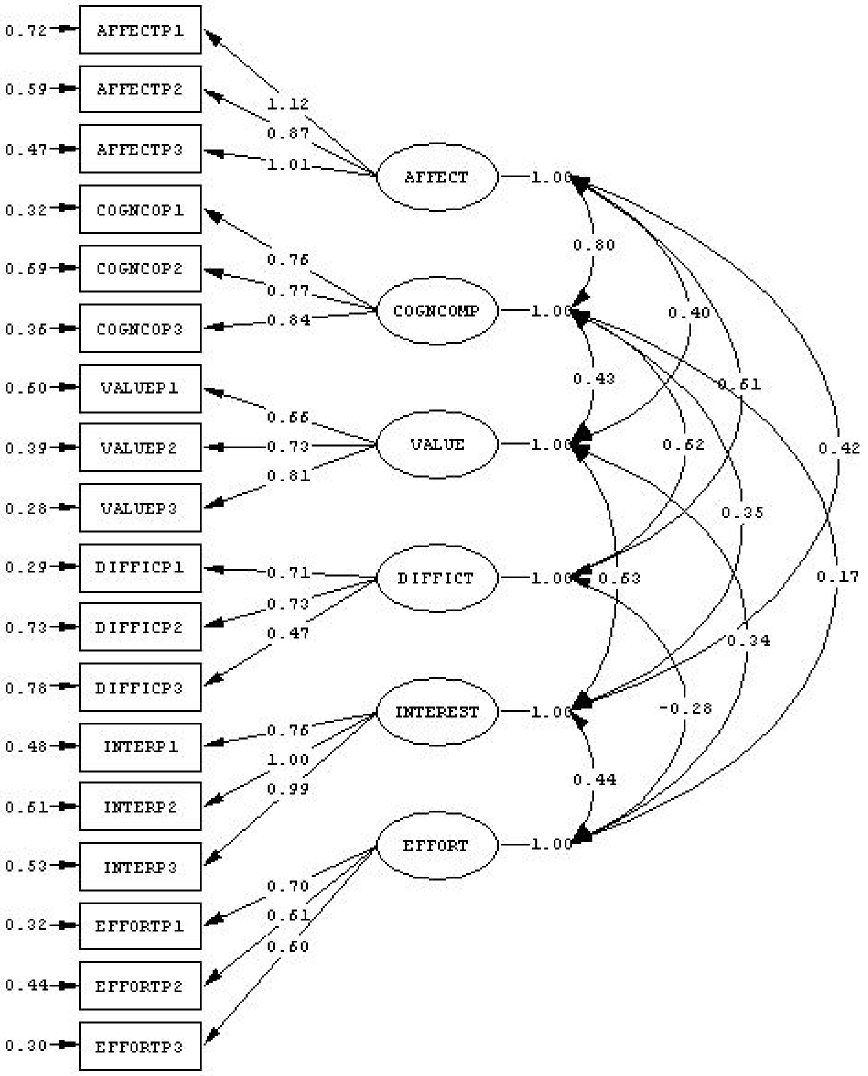


Figure 2. Correlated traits factor model as measurement model for attitudes toward statistics. Values are standardized parameter estimates. All values shown are statistically significant, p < 0.05. AFFECT, Affect; COGNC, Cognitive Competence; VALUE, Value; DIFFIC, Difficulty; INTEREST, interest; EFFORT, (planned) effort.

When interpreting the trait correlation structure, the first issue that comes up is the effect of disentangling the broad task value concept into Affect, related to liking the subject, and Value, related to the importance attached to the subject. The correlation between latent factors Affect and Value (r = 0.40) is, relative to other correlations, modest. This indicates that Affect and Value are clearly empirically distinguishable constructs. The correlation between Value and Difficulty is insignificant, indicating that the attached value is independent to the lack of perceived difficulty. A third observation refers to the by far largest correlation, namely between Affect and Cognitive Competence. This is in itself a remarkable fact: Affect is achieved by decomposing the task value component into affective and utility-related factors, but from this analysis it appears that Affect is much more strongly related to the expectancy component Cognitive

Competence, than to Value. This once again confirms the usefulness of the affect extension of the expectancy-value model. The strong correlation we found is comparable to the results found in Dauphinee et al. (1997) and Hilton et al. (2004).

The relationship between the two factors Interest and Effort and the other four factors is primarily through Value. Interest is unrelated to Difficulty, and Effort is unrelated to Affect and negatively related to Difficulty. That last negative relationship seems to be an consequence of rational study behavior; students who regard statistics as difficult plan to invest more study effort than students regarding the subject as less difficult. However, it is at odds with the expectancy-value model, where that relation has the opposite sign. The different outcome is best explained by the context in which the model is used; whereas the expectancy-value model is primarily based on the selection of learning tasks (such as choosing one course in favor of another), the context of this study is the intensity of performance, given the required learning tasks. In the expectancy-value model, Effort is assumed to be an intermediate outcome variable. For this interpretation to be true, the correlations between Effort and its predictors are expected to be strongly positive. This is not the case, except possibly for Interest. Two potential explanations for the weaker than expected relationship between Effort and its predictors are available. First, Effort is an ex-ante measure, and planned effort might quite well diverge strongly from ex-post measured, realized effort. Second, planned Effort scores seem to be a composition of two rather different underlying mechanisms that can make the relationships of this variable to other attitudinal constructs ambiguous. On the one side, students with high achievement motivation are assumed to spend large efforts in their learning, so planned effort acts as a proxy for achievement motivation. On the other side, planned effort might act as a proxy for students' learning approaches; students with a tendency to a memorizing type of learning tend to invest more effort in their learning than students with a learning approach focused on understanding. In general, the latter deep learning approach is regarded as better, and at least more efficient, than the first mentioned surface learning approach. For that reason, it might be expected that students with a tendency towards deep learning will have more positive attitudes, making deep learning positively related to the several attitudinal variables, and surface learning negatively related. If this is true, the relationship between Effort and attitudinal variables is the result of two counterbalancing forces: higher planned effort levels when being motivated, but lower planned effort levels when relying on efficient, deep learning approaches. In the subsection discussing the outcomes of the full structural equation model, we further elaborate on this issue.

3.4. MEASUREMENT MODEL OF STATISTICAL REASONING ABILITIES

Previous empirical studies of the SRA instrument have used aggregated correct conceptions, and aggregated misconceptions, as scales, with the eight correct reasoning ability scores and the eight misconception scores as items. This would suggest a measurement model with the two aggregated reasoning abilities as latent constructs, and the correct reasoning ability and aggregated variables as indicators. However, Garfield (1998b), Garfield and Chance (2000) and Liu (1998) point out that this modeling approach has important drawbacks. In their studies, as in ours, the correlations between reasoning ability scores are low, mostly insignificant, and quite often of opposite signs. This is problematic in terms of scale construction, because it gives rise to low values of instrument reliability. In the present data set analyzed in this study, the Cronbach-α reliability of the correct reasoning scales is 0.34, whereas for the misconception scales, the reliability α is 0.10. These values are too low to warrant meaningfulness of aggregated constructs. Elsewhere, we have investigated the reliability of aggregated scales for a much larger sample, and have come to similar conclusions (Tempelaar, 2004). Deleting individual items with extreme p-values, as suggested in Liu (1998), appears to have little impact on reliabilities in our data.

Inspection of the correlation matrix depicted in Table 6 does however expose a pattern in correlations that suggests an alternative approach for modelling the outcomes of the SRA-instrument. Correlations within the group of correct reasoning scales, and within the group of misconceptions are, without exception, low. However, in the rectangular part of the correlation matrix containing the correlations between correct reasoning skills and misconceptions, seven out of eight columns contain exactly one highly significant and strongly negative correlation. This is not surprising; from the definition of for example Prob and OutcO, it is apparent that outcome orientation, that is the use of an intuitive and incorrect probability model, is at odds with correctly interpreting probabilities. And in some cases, the strong negative correlations between several correct conceptions and misconceptions find their origin in the fact that the concepts are based on different options of the same multiple choice items, which would lead to negative correlations by construct (although several multiple choice items allow for multiple answers).

Table 6. Correlations between SRA correct reasoning and misconceptions scales being significant at p = 0.01; values in bold exceed 0.30 in absolute value

	Prob	Aver	Comp	Indep	Sampl	Correl	Twow	LrgS
Prob	1.00							
Aver		1.00						
Comp	0.09		1.00					
Indep		0.08	-0.16	1.00				
Sampl		0.10	0.08	-0.07	1.00			
Correl	0.09	0.17				1.00		
Twow	0.13	0.13				0.09	1.00	
LrgS		0.10	0.09		0.07	0.09	0.09	1.00
AverMc		**-0.43**			-0.26			
OutcO	**-0.42**		-0.22	-0.13				**-0.32**
High%						0.08		0.11
Small		-0.10	-0.09	0.07	**-0.69**			-0.16
Repre			-0.21	**-0.69**	0.08			
Cause		-0.08			-0.07	**-0.46**		
EquiPr			**-0.80**	0.20	-0.12	0.09		
Groups								
	AverMc	OutcO	High%	Small	Repre	Cause	EquiPr	Groups
AverMc	1.00							
OutcO		1.00						
High%			1.00					
Small				1.00				
Repre					1.00			
Cause	0.14		-0.07		0.10	1.00		
EquiPr				0.12	-0.10		1.00	
Groups	0.07		0.09					1.00

Taking this pattern of correlations into account, we suggest a different method of aggregating scales scores instead of calculating total correct and misconception scores. On the basis of the strong negative correlations between seven pairs of one correct reasoning scale and one misconception scale, a pair-wise aggregation process seems to be more appropriate than aggregation over all correct, and all incorrect answers. To investigate this option, an exploratory factor analysis was performed. This factor analysis resulted in a seven-factor solution, with five factors composed of pairs of one correct conception and one misconception, having factor loadings of opposite signs: Comp and EquiPr, Sampl and Small, Indep and Repre, Prob and OutcO, and Correl and Cause. The remaining two factors are composed of Aver, Twow, LrgS, and AverMc; and High% and Groups, respectively. All factor loadings have the expected signs: positive for correct conceptions, negative for misconceptions.

Subsequently, a measurement model was estimated taking the outcome of the explorative factor analysis as its basis. No cross-loadings were allowed but, similar to the estimation of the attitudes measurement model, trait correlations were allowed. In addition, uniqueness correlations were allowed for those reasoning abilities and misconceptions that shared an item. Of the 21 trait correlations, only four appear to be significant. This does not come as a surprise, given the many insignificant correlations in Table 6. All 10 uniqueness correlations appear to be significant. The final measurement model for reasoning abilities is depicted in Figure 3; the fit indices of the final model are reported in Table 7. The fit of the CTCU 7 CFA model is good.

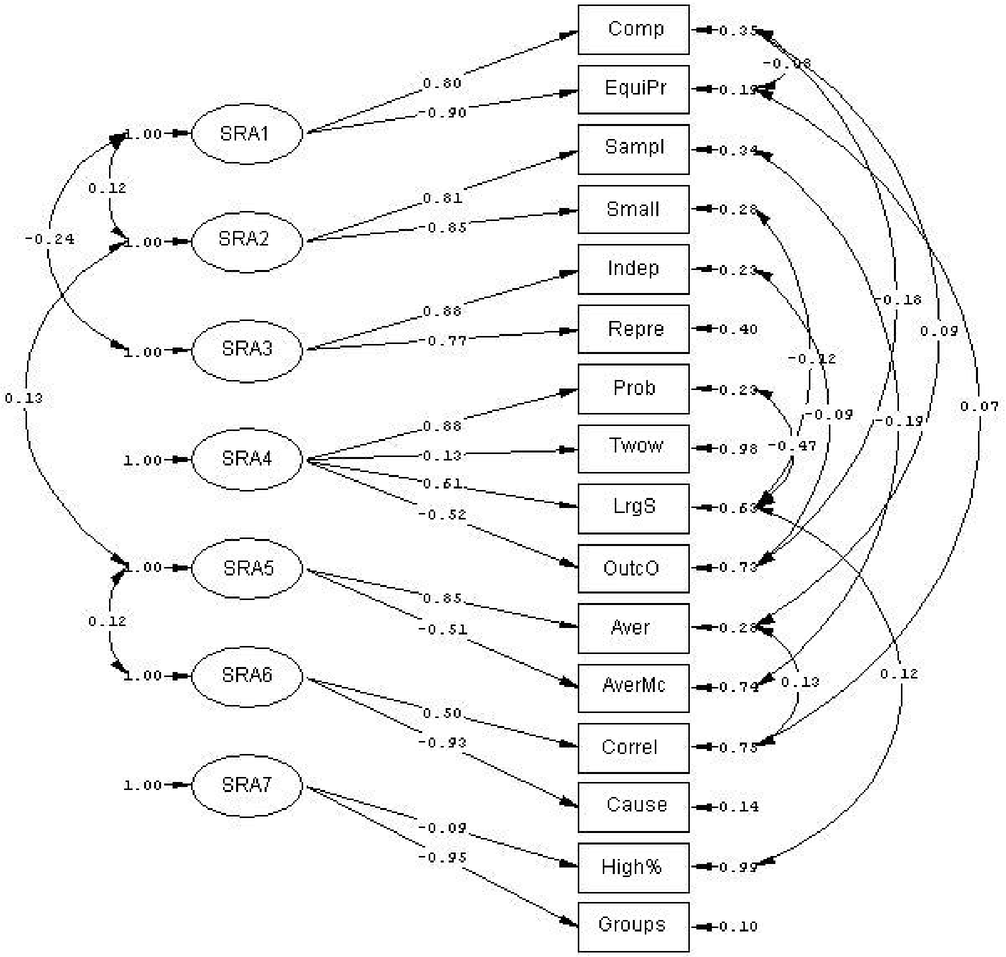


Figure 3. Correlated traits, correlated uniqueness factor model as measurement model for statistical reasoning abilities. Values are standardized parameter estimates. All values shown are statistically significant, p < 0.05. CC, SRA1..7, latent reasoning factors.

Table 7. Fit indices of seven-factor correlated traits confirmatory factor models of statistical reasoning abilities

	χ^2	*df*	RMSEA	GFI	NNFI	CFI	RFI
CTCU 7CFA model	355.00	98	0.042	0.97	0.93	0.94	0. 90

Judging from the good fit of this measurement model, an important conclusion with regard to the SRA instrument becomes apparent. When using SRA as an instrument to assess statistical reasoning, it is less attractive to aggregate all correct scales and all misconception scales into constructs like total correct reasoning and total misconceptions, given the limited reliability of such constructs. As an alternative, composing latent reasoning constructs on which both correct and misconception scales load seems to offer higher reliability.

3.5. FULL STRUCTURAL EQUATION MODEL OF ATTITUDE AND BELIEFS, STATISTICAL REASONING ABILITIES, AND COURSE PERFORMANCE

The final step in the analysis regards the integration of both measurement models. This includes the not explicitly elaborated model for course performances, specifying the two latent course performances EXAM and QUIZ. Both course performance constructs are measured by two indicators: a score for mathematics and a score for statistics. The relationships that link the latent factors in the three measurement parts constitute the structural part of the model. The estimation of the structural parameters is similar to the estimation of trait correlations in the measurement models; no a priori restrictions apply as to what parameters are restricted to zero and which are set free. Two modification directions were applied: model building and model trimming. Both methods converge to the model depicted in Figure 4. Figure 4 does not make explicit the estimated correlations between latent factors; the same correlation structure as visible in Figures 2 and 3 was however used in the estimation of the full model. Table 8 reports fit indices of that model and indicates good fit. Table 9 describes the standardized parameter estimates or β-coefficients of the structural part of the model.

Table 8. Fit indices of full structural model of attitudes toward statistics, statistical reasoning abilities, and course performance

	χ^2	*df*	RMSEA	GFI	NNFI	CFI	RFI
SEM	1599.26	620	0.035	0.94	0.96	0.97	0.94

Table 9. Standardized estimates of the structural part of the full structural model of attitudes toward statistics, statistical reasoning abilities, and course performance

	Affect	Cog Comp	Value	Difficulty	Interest	Effort	SRA4
SRA1						-0.10	
SRA2				0.09		-0.09	
SRA3							
SRA4			0.08			-0.10	
SRA5				0.18			
SRA6				0.10	-0.08		
SRA7			0.07			-0.12	
EXAM		0.39		-0.17			0.56
QUIZ		0.34		-0.17		0.09	

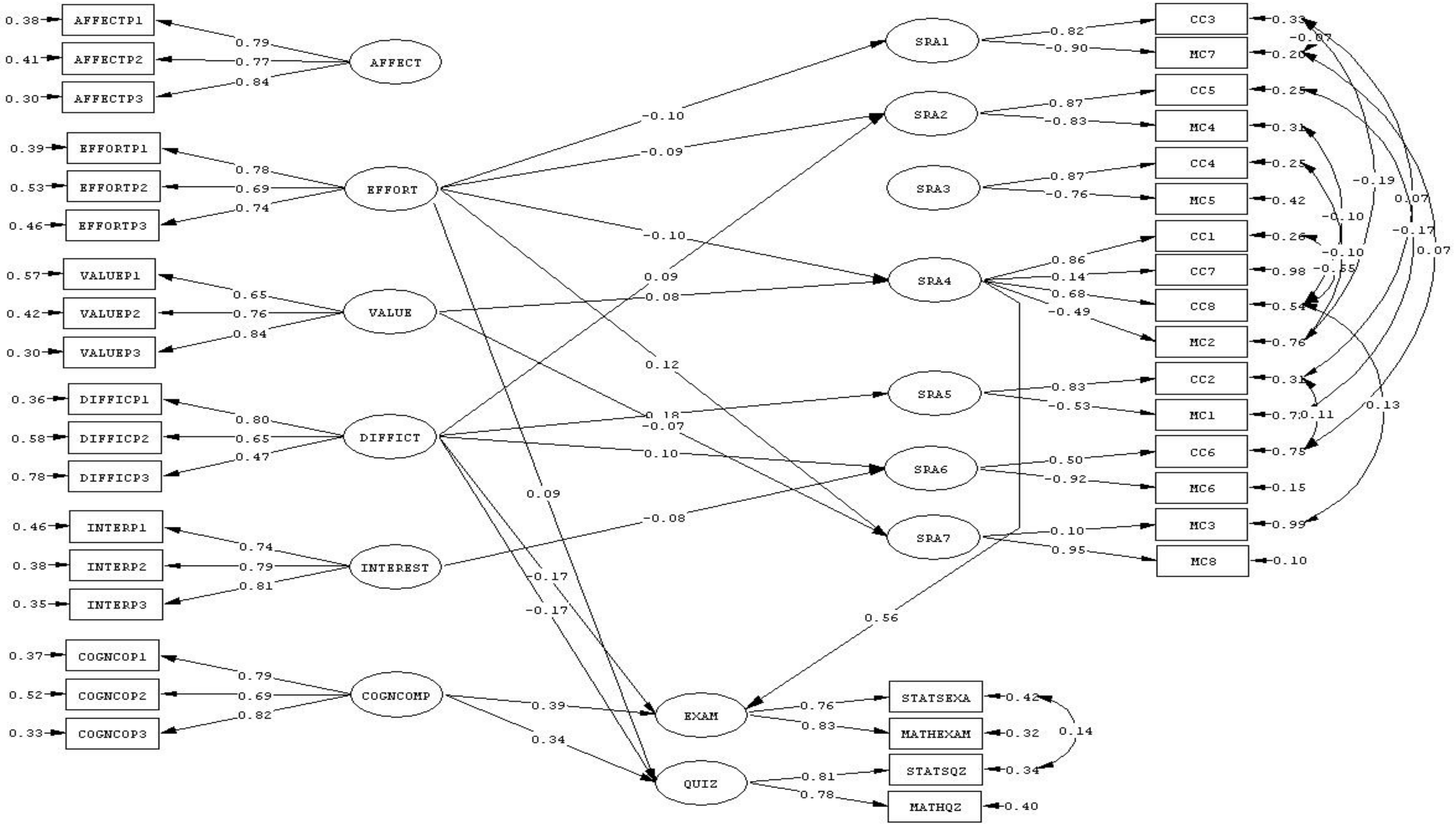


Figure 4. *Structural equation model of attitudes toward statistics, statistical reasoning abilities, and course performance. Values are standardized parameter estimates. All values shown are statistically significant, p < 0.05. AFFECT, Affect; COGNC, Cognitive Competence; VALUE, Value; DIFFIC, Difficulty; INTEREST, interest; EFFORT, (planned) effort, SRA1..7, latent reasoning factors; EXAM, QUIZ, latent course performance factors*

The final structural model allows several interpretations. Students' self-ability belief, Cognitive Competence, is a strong predictor of both latent course performance factors, with β-values of 0.39 and 0.34. This is in agreement with many studies on the expectancy-value model, and self-concept or self-efficacy research. The relationships between statistical reasoning and the two course performance factors are weak, which is in line with the low correlations between SRA constructs and course performance found in several studies. In our study, only SRA4, the latent factor composed of four correct conceptions and misconceptions related to the ability to interpret probabilities, has a significant and strong impact on the latent exam factor.

The second direct effect from attitudinal variables on course performances stems from the other expectancy construct of perceived task demand: (lack of) Difficulty. The relationship is reversed, with β-values of -0.17. This outcome is somewhat surprising; the expectancy-value model would predict a positive relationship. However, the relationship is robust; using a split-sample approach (and path analysis), it is confirmed in subsamples composed in several ways. The bivariate relationship between Difficulty and performance is however absent; the negative relation we find is present only in a simultaneous relation between Cognitive Competence, Difficulty, and course performance. It should thus be interpreted as a process of underestimation of task demand by students with an above average ability belief.

The reduced form squared multiple correlations of both course performance latent factors EXAM and QUIZ are equal to 0.10. This means that the combined effect of both direct paths from SATS variables to EXAM and QUIZ, and the indirect paths from SATS via SRA to the two course performance factors, explains 10% of the total variation in both course performances. In the decomposition of explained variation into direct and indirect effects, it becomes clear that the contribution of the indirect effect can be ignored: less than 0.5%. The dominance of direct over indirect effects is due to the fact that relations between SATS and SRA are weak, and much weaker than relations between SATS and performance. In line with the expectancy-value model, attitudes have a positive impact on reasoning abilities through the variables Value and (perceived lack of) Difficulty. In contrast to predictions based on the expectancy-value model, the Effort variable has a negative impact on four of the seven latent reasoning factors. The negative relationship is consistent: β-coefficients of Effort to the several SRAs are either significantly negative, or zero, but never positive. Although a negative relation may appear counter-intuitive, it is in line with related research on the relationship between preferred learning approaches and reasoning abilities, where it was found that a tendency to surface learning negatively influences statistical reasoning (Tempelaar, 2004; Tempelaar, Gijselaers, & Schim van der Loeff, 2006; Tempelaar, Schim van der Loeff, Gijselaers, Crombrugghe, 2007). Planned effort being a proxy of both achievement motivation and a non-efficient learning approach (see the above discussion of the measurement model of attitudes), will give rise to diverse relationships between learning outcomes and the Effort variable. Learning performances that allow for alternative learning paths – such as memorizing versus understanding – are expected to demonstrate a positive relationship with planned effort. For these learning performances, the achievement motivation component in planned effort is dominant; students who are prepared to work hard will achieve better performances. In our study, quiz scores for both mathematics and statistics are the ultimate example of such type of course performances. Quizzes are designed to be accessible for all students and the bonus points they bring about are especially helpful for students at risk of not passing the course. This makes it plausible that the motivation component in planned effort dominates the learning approach component, which explains the positive relationship between Effort and Quiz.

The opposite case is constituted by the SRA factors. Because the SRA is administered as an entry measurement unrelated to course grading, any direct effect of achievement motivation can assumed to be absent. And because statistical reasoning is not part of any secondary education of most students in this study, indirect effects – taking advantage of having been highly motivated in secondary school – will at most be very modest. As a result, the learning approach component in planned effort is expected to be dominant, which quite well explains the negative relationships found between EFFORT and four of the SRA factors. In this spectrum of course performances, the scores on the exam take an intermediate position. Being the course performance measurement, they certainly contain a strong achievement motivation component. At the same time, exams are certainly much less accessible than quizzes, which feeds the learning approach component. In the aggregation, the two effects are counterbalancing, which quite well might explain the latent factor EXAM being unrelated to Effort.

4. CONCLUSIONS

In this study the affect-extended version of the expectancy-value model (Schau et al., 1995; Dauphinee et al., 1997; Hilton et al., 2004) was adopted as an achievement motivation model. Our data corroborate this extension, in the sense that affect and value turn out to be clearly distinguishable constructs, as well as in the sense that these variables play a distinctive role in the relationships with reasoning abilities and course performance. To our knowledge this study is the first to apply the 36-item SATS version, with the new scales Interest and Effort. Both scales appear to be a valuable addition to the instrument. The latent trait correlations in Table 5 demonstrate that the two factors are well identified constructs. However, correlational analysis suggests that Effort might be composed of two rather different characteristics. Therefore, a decomposition of this scale into an achievement motivation aspect and a learning approach aspect is called for. The latter aspect has the interpretation that students with a surface learning approach will typically achieve high scores on this Effort variable, because they invest large amounts of time for learning subjects by memorization.

Through a factor-analytic study, we conclude that a factor model with most factors being composed of pairs of one reasoning ability and one misconception provides an appropriate measurement model. This shows that the SRA-instrument used by Garfield (1998b, 2003), Garfield and Chance (2000) and Liu (1998) is not flawed. In studies by these authors only two aggregate scales, one for statistical reasoning abilities and one for statistical misconceptions, are employed. They point out that these aggregate scales have shortcomings in view of the low values of correlations between the scales that constitute both aggregate scales, which results in low reliabilities. Our results imply that the finding of low correlations does not invalidate the instrument, but that alternative measurement models other than the one based on aggregate scales should be used.

This study adds support to previous findings of the absence of a strong relationship of misconceptions and their counterpart, the reasoning abilities, with students' course performances. This is demonstrated in studies where statistical reasoning is regarded as one of the several learning outcomes of the course and assessed simultaneously with these other course performances (Garfield, 1998b, 2003; Garfield and Chance, 2000; and Liu, 1998). In the present study along with those of Tempelaar (2004) and Tempelaar et al. (2006), it is also demonstrated in a second type of studies, where statistical reasoning is regarded as part of the prior knowledge state of the student and assessed before the start of the course. Are these studies, given their conclusions that SRA components are only weakly or even un-related to different course performance indicators, uninformative? We would argue that the opposite is true; exactly because of these absent relationships, they are informative. In general, different components of statistical knowledge, measured as course performance scores, tend to be substantially correlated. For example, in this study the correlation between latent course performance factors EXAM and QUIZ equals r = 0.69. And investigating the relationships among three rather different types of course performances, final exam scores, quiz scores, and homework scores, we find similar substantial correlations. Because the SRA-instrument was developed to assess statistical reasoning mastery achieved in high school statistics programs, the natural hypothesis is that SRA-scores correlate with the several course performances in the same way as the other components of course performances do. But they clearly do not do so. It is these unexpected low correlations that make studies such as ours informative, rather than the case that the expected, substantial positive correlations would have been found.

The absence of substantial relationships can be well explained in the context of naïve theories that are an element of the new theory of learning, as elaborated in Bruer's (1993) 'Schools for thought.' Naïve theories or misconceptions are informal, self-acquired elements of science knowledge, inconsistent with formal science. Students can possess formal knowledge and naïve knowledge at the same time; the learning of formal knowledge does not automatically imply that naïve knowledge is unlearned. In spite of having mastered the formal knowledge, students tend to solve scientific problems with their naïve knowledge, especially when they are confronted with these problems outside a school context. And, worst of all, formal knowledge tends to be forgotten much faster than naïve knowledge. Empirical outcomes of studies using the SRA-instrument are in line with these observations. Absence of substantive relationships is compatible with the hypothesis that both statistical reasoning abilities and statistical misconceptions are part of students' naïve statistical knowledge; the first category naïve and correct, the second category naïve but incorrect. More research to investigate the role of naïve theories in learning and the development of naïve knowledge over time is necessary. This is particularly relevant because the reform movement in statistics education has called for a more prominent position of statistical reasoning, and the related domains of statistical literacy and thinking in the statistics curriculum. So it is the reformed curriculum, more than any traditional curriculum, that requires resolving the instructional challenge of unlearning statistical misconceptions before being able to replace them with proper reasoning abilities.

Empirical studies as documented in special issues of *SERJ* (Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005) and in Ben-Zvi & Garfield (2005) conclude that in order to learn reasoning and to unlearn misconceptions, the use of specific educational tools is indispensable. This study suggests that the use of these tools is probably only part of the solution of the instructional challenge. A strong dependency on these instructional tools might be at odds with educational principles on which student-centered programs are based, in the sense that they limit students' own responsibility to organize the learning process. The outcomes of this study might bring forward some further limitations. In most learning processes students enter the learning context with a given set of background characteristics, such as a preference for deep learning versus surface learning. Most of these contexts allow all students to achieve satisfactory learning outcomes, be it along different learning paths. As a concrete example, our structural equation model suggests that both surface learning oriented students and deep learning oriented students can achieve adequate course performance scores. But our empirical analyses also suggest that statistical reasoning might be the odd man out in this context; the learning of statistical reasoning seems not easily to assimilate to the variation in students' background characteristics as preferred learning approach, as is the case with other cognitive goals. If this conclusion is correct, it implies we need an even broader range of educational tools than already described in the sources referred to earlier; more than content, the tools should address general learning approaches.

ACKNOWLEDGEMENTS

The authors express their gratitude to the anonymous referees and the assistant editor, Beth Chance, for their helpful remarks and suggestions.

REFERENCES

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004a). *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, The Netherlands: Kluwer Academic Publishing.

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004b). Research on reasoning about variability [Special issue]. *Statistics Education Research Journal, 3*(2).

Ben-Zvi, D., & Garfield, J. B. (2004c). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bruer, J. T. (1993). *Schools for thought: A science of learning in the classroom*. Cambridge, MA: The MIT Press.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics scale: A construct validity study. *Educational and Psychological Measurement*, *65*(3), 509-524.

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

Chance, B. L. & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal*, *1*(2), 38-41.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ1(2).pdf]

Dauphinee, T. L., Schau, C., & Stevens, J. J. (1997). Survey of Attitudes Toward Statistics: Factor structure and factorial invariance for women and men. *Structural Equation Modeling*, *4*(2), 129-141.

delMas, R. (2002). Statistical literacy, reasoning, and learning. *Journal of Statistics Education*, *10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/delmas_intro.html and http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html]

delMas, R. (2004a).A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishing.

delMas, R. (2004b). Overview of ARTIST website and Assessment Builder. *Proceedings of the ARTIST Roundtable Conference*, Lawrence University.
[Online: http://www.rossmanchance.com/artist/Proctoc.html]

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105-121). New York: The Guilford Press.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109-132.

Gal, I. (2004). Statistical literacy, meanings, components, responsibilities. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 47-78). Dordrecht, The Netherlands: Kluwer Academic Publishing.

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

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2).
[Online: http://www.amstat.org/publications/jse/v2n2/gal.html]

Garfield, J. B. (1996). Assessing student learning in the context of evaluating a chance course. *Communications in Statistics; Part A: Theory and Methods*, *25*, 2863-2873.

Garfield, J. B. (1998a, April). *Challenges in assessing statistical reasoning*. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22-38.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(1).pdf]

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

Garfield, J. B., & Ben-Zvi, D. (2004a). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Garfield, J. B., & Ben-Zvi, D. (2004b). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 397-409). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Garfield, J. B., & Ben-Zvi, D. (Eds.) (2005). Reasoning about variation [Special section]. *Statistics Education Research Journal*, *4*(1).

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

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

Harris, M. B., & Schau, C. (1999). Successful strategies for teaching statistics. In S.N. Davis, M. Crawford, & J. Sebrechts (Eds.), *Coming into her own: Educational success in girls and women* (pp. 193-210). San Francisco: Jossey-Bass.

Hau, K. T., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology*, *57*, 327-351.

Hilton, S. C., Schau, C., & Olsen, J. A. (2004). Survey of Attitudes Toward Statistics: Factor structure invariance by gender and by administration time. *Structural Equation Modeling*, *11*(1), 92-109.

Jolliffe, F. (1998). What is research in statistical education? In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 801-806). Voorburg, The Netherlands: International Statistical Institute.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, MA: Cambridge University Press.

Kline, R. B. (2005). *Principles and practice of structural equation modelling* (2nd ed.). New York: Guilford Press.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*, 59-98.

Liu, H. J. (1998). *A cross-cultural study of sex-differences in statistical reasoning for college students in Taiwan and the United States*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220.

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning, a project of the National Council of Teachers of Mathematics* (pp. 575-596). New York: Macmillan.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishing.

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

Schau, C. (2003, August). *Students' attitudes: The "other" important outcome in statistics education*. Paper presented at the Joint Statistical Meetings, San Francisco.

Schau, C., Dauphinee, T. L., Del Vecchio, A., & Stevens, J. (1999). *Survey of attitudes toward statistics (SATS).*
[Online: http://www.unm.edu/~cschau/downloadsats.pdf]

Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement, 55*(5), 868-875.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

Short, T. H. (Ed.) (2002). Statistical literacy, reasoning, and thinking [Special section]. *Journal of Statistics Education, 10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/abstracts.html]

Sorge, C., & Schau, C. (2002, April). *Impact of engineering students' attitudes on achievement in statistics*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Sundre, D. L. (2003, April), *Assessment of Quantitative reasoning to enhance educational quality*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
[Online: http://www.gen.umn.edu/artist/articles/AERA_2003_QRQ.pdf]

Tempelaar, D. (2004). Statistical reasoning assessment: An Analysis of the SRA instrument. *Proceedings of the ARTIST Roundtable Conference*, Lawrence University.
[Online: http://www.rossmanchance.com/artist/Proctoc.html]

Tempelaar, D. T., Gijselaers, W. H., & Schim van der Loeff, S. (2006). Puzzles in statistical reasoning. *Journal of Statistics Education, 14*(1). [Online: http://www.amstat.org/publications/jse/v14n1/tempelaar.html]

Tempelaar, D. T., Gijselaers, W.H., Schim van der Loeff, S., & Nijhuis, J. (2007). A structural equation model analyzing the relationship of student achievement motivations and personality factors in a range of academic subject-matter areas. *Contemporary Educational Psychology, 32*(1), 105-131.

Tempelaar, D., Schim van der Loeff, S., Gijselaers, W., & De Crombrugghe, D. (2007). *Preferred learning approaches and statistical reasoning.* Unpublished manuscript.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68-81.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 92-122). San Diego: Academic Press.

DIRK T. TEMPELAAR
Department of Quantitative Economics,
Faculty of Economics and Business Administration
University of Maastricht
PO Box 616, 6200 MD Maastricht
the Netherlands

PAST IASE CONFERENCES

SRTL-5
THE FIFTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING, AND LITERACY
Coventry, UK, August 11 - 17, 2007

Reasoning about Statistical Inference: Innovative Ways of Connecting Chance and Data.

The fifth research forum in a series of international research forums on statistical reasoning, thinking and literacy was held at the Centre for New Technologies Research in Education of the University of Warwick, England. This particular gathering of researchers played an important role in advancing our understanding of the richness and depth of reasoning about informal inference, a natural development of previous foci on variability and distribution. The forum was sponsored by The Royal Statistical Society (UK), The American Statistical Association (ASA) Section on Statistical Education, the Institute of Education, University of Warwick, and the School of Education, University of Leicester.

Twenty-four researchers in statistics education from seven countries shared their work, discussed important issues, and initiated collaborative projects in a stimulating and enriching environment. Sessions were held in an informal style, with a high level of interaction. With emphasis on reasoning about informal inference, a wide range of research projects were presented spanning learners of all ages, as well as teachers and practitioners in the workplace. These demonstrated an interesting diversity in research methods, theoretical approaches, and points of view. As a result of the success of this gathering, plans are already underway for the next gathering (SRTL-6) in 2009.

The research forum proved to be very productive in many ways. Progress was made towards identifying the key elements of statistical inference and in locating the range of resources that might be brought to bear in supporting engagement with those powerful ideas. Several types of scientific publications will be produced including proceedings on the Forum Website (http://srtl.stat.auckland.ac.nz/), papers in refereed journals, and a special issue of *Statistics Education Research Journal*, expected in 2008, with Dave Pratt and Janet Ainley as guest editors. These outcomes will all serve as a rich resource for statistics educators and researchers.

Dave Pratt (Institute of Education, University of London) and Janet Ainley (University of Leicester) led the local planning and organizing prior to the SRTL-5 gathering. Yvette Kingston, supported by Peter Johnston-Wilder and Theodosia Prodromou (all University of Warwick) ensured that the forum ran smoothly and was able to meet its objectives. Thanks to the efforts of this group, participants were able to not only enjoy each other’s creative efforts during the scientific programme but also to appreciate the local culture through a variety of social events that helped to build a sense of a community amongst the researchers.

For further information please contact the SRTL co-chairs:

Joan Garfield, jbg@umn.edu and Dani Ben-Zvi, dbenzvi@univ.haifa.ac.il

IASE SATELLITE CONFERENCE ON ASSESSING STUDENT LEARNING IN STATISTICS Guimarães, Portugal, August 19-21, 2007

The meeting was held on August 19-21, 2007 in Guimarães, Portugal, immediately prior to ISI-56 in Lisbon. The Satellite involved papers on many aspects of assessing student learning in statistics. Over 40 papers were presented along with a number of posters and a discussion of examination questions.

The complete proceedings, published also on CD, is freely available at

http://www.swinburne.edu.au/lss/statistics/IASE/CD_Assessment/index.htm

Papers presented at the conference are also available at

http://www.stat.auckland.ac.nz/~iase/publications.php?show=sat07

ISI-56 THE 2007 SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE Lisboa, Portugal, August 22 – 29, 2007

The 56th Session of the International Statistical Institute (ISI) was held in Lisboa, Portugal. The International Association for Statistical Education (IASE) organized 10 statistics education sessions for ISI-56. It is planned that the papers presented in the IASE sponsored sessions will be available at the website:

http://www.stat.auckland.ac.nz/~iase/publications.php

FORTHCOMING IASE CONFERENCES

JOINT ICMI /IASE STUDY
STATISTICS EDUCATION IN SCHOOL MATHEMATICS: CHALLENGES FOR TEACHING AND TEACHER EDUCATION
Monterrey, Mexico, June 30 to July 4, 2008

The International Commission on Mathematical Instruction (ICMI, http://www.mathunion.org/ICMI/) and the International Association for Statistical Education (IASE, http://www.stat.auckland.ac.nz/~iase/) are pleased to announce the Joint ICMI /IASE Study - Statistics Education in School Mathematics: Challenges for Teaching and Teacher Education.

Following the tradition of ICMI Studies, this Study will comprise two parts: the Joint Study Conference and the production of the Joint Study book. The Joint Study Conference will be merged with the IASE 2008 Round Table Conference.

The Joint Study Conference (ICMI Study and IASE Round Table Conference) will take place at the Instituto Tecnológico y de Estudios Superiores, Monterrey, Mexico (http://www.mty.itesm.mx/), from June 30 to July 4, 2008. Participation in the Conference is only by invitation, based on a submitted contribution and a refereeing process. Accepted papers will be presented in the Conference and will appear in the Proceedings that will be published by ICMI and IASE as a CD-ROM and on the Internet.

The second part of the Joint Study – the Joint Study book – will be produced after the conference and will be published in the ICMI Study Series. Participation in the Joint Study Conference does not automatically assure participation in the book, since a second selection and rewriting of selected papers will be made after the conference.

More information: Carmen Batanero, batanero@ugr.es

Website: http://www.stat.auckland.ac.nz/~iase/temp/RoundTable2008Announce.htm

ICME-11
INTERNATIONAL CONGRESS ON MATHEMATICAL EDUCATION
TOPIC STUDY GROUP #13
RESEARCH AND DEVELOPMENT IN THE TEACHING AND LEARNING OF PROBABILITY
Monterrey, Mexico – July 6 - 13, 2008

Probability and statistics education are relatively new disciplines. Both have only recently been introduced into main stream school curricula in many countries. While application-oriented statistics is undisputed in its relevance, discussion about probability is more ambivalent. When probability is reduced to its classical conception, mainly based on combinatorics or its formal treatment in higher mathematics, it can be seen as irrelevant, and may be abandoned to leave only the statistical element of the stochastics discipline. However, we believe that there are some powerful arguments in favour of a strong role for probability within stochastics curricula.

We invite submissions related to the following topics:

Individuals' corner

- Students' understanding and misunderstanding of fundamental probabilistic concepts
- Ideas of probability in young children

Impact of technology

- The use of technology for students' learning of probability
- Using specific software to study probability and sampling distributions
- Special issues in e-learning

Teacher's corner

- Teacher education on the topic of probability
- Teachers' conceptions about teaching probability

Fundamental ideas

- The probabilistic idea of random variable; distribution, expectation
- The central limit theorem; convergence
- Bayes' theorem and conditional probability; independence; exchangeability
- Probabilistic modelling – a probabilistic look at distributions

TEAM CHAIRS

Manfred Borovcnik (Austria), manfred.borovcnik@uni-klu.ac.at
Dave Pratt (U.K.), d.pratt@ioe.ac.uk
Silvia Alatorre Frenk (Mexico), alatorre@solar.sar.net

TEAM MEMBERS

Carmen Batanero (Spain), batanero@ugr.es
Wu Yingkang (China) , ykwu@math.ecnu.edu.cn

Website: http://tsg.icme11.org/tsg/show/14

ICME-11
INTERNATIONAL CONGRESS ON MATHEMATICAL EDUCATION
TOPIC STUDY GROUP #14
RESEARCH AND DEVELOPMENT IN THE TEACHING AND LEARNING OF STATISTICS
Monterrey, Mexico – July 6 - 13, 2008

Statistics education is a growing field of research and development at school and university level. The topic group will focus on presenting and discussing recent research.

Statistics at school level is usually taught in the mathematics classroom in connection with learning probability. Inferential statistics is based on basic understandings of probability. Our topic includes probabilistic aspects in learning statistics, whereas research with a specific focus on learning probability is being discussed Topic Study Group #13 of ICME.

We are open to all kinds of relevant research papers, but our specific focus will be on the following topics:

- Students' thinking and reasoning about distributions (including variability, comparing distributions)

- Students’ making inferences from data (from informal inference to more formal inference, inference from sample to population or process, from data to context, role of models and probability)
- Statistical literacy
- Role of technology (tools, applets, internet)
- Research on teachers and teaching of statistics

TEAM CHAIRS

Rolf Biehler (Germany), biehler@mathematik.uni-kassel.de
Mike Shaughnessy (USA), mikesh@pdx.edu

TEAM MEMBERS

Omar Rouan (Morocco), orouan@yahoo.com
Ernesto Sánchez (Mexico), esanchez@cinvestav.mx
Jane Watson (Australia), Jane.Watson@utas.edu.au

Website: http://tsg.icme11.org/tsg/show/15

ISI-57
THE 2009 SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE
Durban, South Africa, August 16 – 22, 2009

IASE sponsored Invited Paper Meetings for 57th Session in Durban are being organised by Helen MacGillivray (Australia, h.macgillivray@qut.edu.au). The IASE Programme Committee for ISI-57 has chosen the theme - Statistics Education for the Future.

More information is available at:
http://www.statssa.gov.za/isi2009/

SRTL-6
THE SIXTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING, AND LITERACY
Brisbane, Australia, 2009

The sixth SRTL forum will be organized at University of Queensland, Brisbane.

The forum coordinator is Katie Makar (k.makar@uq.edu.au).

ICOTS-8
DATA AND CONTEXT IN STATISTICS EDUCATION: TOWARDS AN EVIDENCE-BASED SOCIETY
Ljubljana, Slovenia, July 11-16, 2010

The 2010 International Conference on Teaching Statistics will be held in the city of Ljubljana, Slovenia, July 11-16. It is being organised by the IASE and the Slovenian Statistical Association. The venue will be the Ljubljana Cultural and Congress Centre.

Statistics educators, statisticians, teachers and educators at large are invited to contribute to the scientific programme. Types of contribution include invited papers, contributed papers, and posters. No person may author more than one Invited Paper at the conference, although the same person can be co-author of more than one paper, provided each paper is presented by a different person.

Voluntary refereeing procedures will be implemented for ICOTS-8. Details of how to prepare manuscripts, the refereeing process and final submission arrangements will be announced later.

INVITED PAPERS

Invited Paper Sessions are organized within 10 Conference Topics as follows.

Topics and Topic Convenors

1. Data and Context in Statistics Education: Towards an Evidence-based Society.
   Brian Phillips (Australia) bphillips@swin.edu.au
   Irena Ograjensek (Slovenia) irena.ograjensek@ef.uni-lj.si
2. Statistics Education at the School Level.
   Mike Shaughnessy (USA) mikesh@pdx.edu
   Doreen Connor (UK) doreen.connor@ntu.ac.uk
3. Learning to Teach Statistics.
   Katie Makar (Australia) k.makar@uq.edu.au
   Joachim Engel (Germany) engel@math.uni-hannover.de
4. Statistics Education at the Post Secondary Level.
   Elisabeth Svensson (Sweden) elisabeth.svensson@esi.oru.se
   Larry Weldon (Canada) weldon@sfu.ca
5. Assessment in Statistics Education.
   Beth Chance (USA) bchance@calpoly.edu
   Iddo Gal (Israel) iddo@research.haifa.ac.il
6. Statistics Education, Training and the Workplace.
   Gabriella Belli (USA) gbelli@vt.edu
   Peter Petocz (Australia) peter.petocz@mq.edu.au
7. Statistics Education and the Wider Society.
   Richard Gadsden (UK) R.J.Gadsden@lboro.ac.uk
   Oded Meyer (USA) meyer@stat.cmu.edu
8. Research in Statistics Education.
   Arthur Bakker (The Netherlands) a.bakker@fi.uu.nl
   Tim Burgess (New Zealand) t.a.burgess@massey.ac.nz
9. Technology in Statistics Education.

Deborah Nolan (USA) nolan@stat.berkeley.edu
Paul Darius (Belgium) paul.darius@biw.kuleuven.be

10. An International Perspective on Statistics Education.
Delia North (South Africa) northd@ukzn.ac.za
Enriqueta Reston (Phillipines) edreston@usc.edu.ph

Session themes within each Topic are currently being discussed. The themes and Session organizers with email contact will be available on the ICOTS-8 web site http://icots8.org/, under “Scientific Programme” by June 2008. Those interested in submitting an invited paper should contact the appropriate Session Organiser before December 1, 2008.

CONTRIBUTED PAPERS

Contributed paper sessions will be arranged in a variety of areas. Those interested in submitting a contributed paper should contact either Gilberte Schuyten (Gilberte.Schuyten@UGent.be), John McKenzie (mckenzie@babson.edu), or Flavia Jolliffe (F.Jolliffe@kent.ac.uk) before September 1, 2009.

POSTERS

Those interested in submitting a poster should contact Mojca Bavdaz (mojca.bavdaz@ef.uni-lj.si) or Alesa Lotric Dolinar (alesa.lotric.dolinar@ef.uni-lj.si) before January 15, 2010.

GENERAL ISSUES

More information is available from the ICOTS-8 web site at http://icots8.org/ which will continue to be updated over the next three years, or from the ICOTS IPC Chair John Harraway, (jharraway@maths.otago.ac.nz), the Programme Chair Roxy Peck (rpeck@calpoly.edu), and the Scientific Secretary Helen MacGillivray (h.macgillivray@qut.edu.au).

OTHER PAST CONFERENCES

USCOTS 2007
UNITED STATES CONFERENCE ON TEACHING STATISTICS
Columbus OH, USA, May 17-19, 2007

The second biennial United States Conference on Teaching Statistics (USCOTS 07) was held on May 17-19, 2007 at the Ohio State University in Columbus, Ohio, hosted by CAUSE, the Consortium for the Advancement of Undergraduate Statistics Education. The target audience for USCOTS was teachers of undergraduate and AP statistics, from any discipline or type of institution.

Materials for most of the talks are available at USCOTS page:
http://www.causeweb.org/uscots/program/

2007 JOINT STATISTICAL MEETINGS
Salt Lake City UT, USA, July 29 - August 2, 2007

JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Some materials are available at http://www.amstat.org/meetings/jsm/2007/.

JOINT SOCR (STATISTICS ONLINE COMPUTATIONAL RESOURCE) CAUSEWAY CONTINUING EDUCATION WORKSHOP 2007
UCLA, Los Angeles CA, USA, August 6-8, 2007

The 2007 joint SOCR/CAUSEway continuing education workshop aimed at demonstrating the functionality, utilization and assessment of the current UCLA, SOCR and CAUSEweb resources. This workshop appealed to AP teachers and college instructors of probability and statistics classes who have interests in exploring novel IT-based approaches for enhancing statistics education.

Workshop materials, including the Workshop Handbook, are freely available at
http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_SOCR_CAUSEway_Workshop2007

9TH INTERNATIONAL CONFERENCE OF THE MATHEMATICS EDUCATION INTO THE 21ST CENTURY PROJECT
MATHEMATICS EDUCATION IN A GLOBAL COMMUNITY
Charlotte NC, USA, September 7 - 13, 2007

The Mathematics Education into the 21st Century Project was founded in 1986 and is dedicated to the planning, writing and disseminating of innovative ideas and materials in Mathematics and Statistics Education. Conference materials and presented papers are available at: math.unipa.it/~grim/21_project/21_charlotte_2007.htm

OTHER FORTHCOMING CONFERENCES

INTED 2008
INTERNATIONAL TECHNOLOGY, EDUCATION AND DEVELOPMENT CONFERENCE
Valencia, Spain, March 3 – 5, 2008

On behalf of the INTED 2008 Organizing Committee we would like to invite you to participate in the International Technology, Education and Development Conference in Valencia (Spain) on the 3rd, 4th and 5th of March, 2008.

The general objective of INTED 2008 conference is the promotion of international collaboration in the field of technology, engineering, and science education. INTED 2008 provides an International Forum for researchers, engineers, professors, educational scientists and technologists in the areas of Education, Science, and Technology. It will be an excellent opportunity to present, demonstrate and discuss research, development, applications, and the latest innovations and results in the field of Higher Education and Industry.

More information: inted2008@iated.org

Website: http://www.iated.org/inted2008/

2008 JOINT STATISTICAL MEETINGS
Denver CO, USA, August 3 - 7, 2008

JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Attended by over 5000 people, activities of the meeting include oral presentations, panel sessions, poster presentations, continuing education courses, exhibit hall (with state-of-the-art statistical products and opportunities), career placement service, society and section business meetings, committee meetings, social activities, and networking opportunities. Denver, the host city for JSM 2008, offers a wide range of possibilities for sharing time with friends and colleagues.

More information: jsm@amstat.org

Website: http://www.amstat.org/meetings/jsm/2008/

10TH INTERNATIONAL CONFERENCE OF THE MATHEMATICS EDUCATION INTO THE 21ST CENTURY PROJECT
MODELS IN DEVELOPING MATHEMATICS EDUCATION
Dresden, Saxony, Germany, September 11 – 17, 2009

The Mathematics Education into the 21st Century Project was founded in 1986 and is dedicated to the planning, writing and disseminating of innovative ideas and materials in Mathematics and Statistics Education. You are invited to attend our 10th anniversary project conference to be held in the historic city of Dresden, Germany. The chairman of the Local Organising Committee will be Prof. Dr. Ludwig Paditz of the Dresden University of Applied Sciences.

More information: Alan Rogerson, arogerson@inetia.pl

Website: http://math.unipa.it/~grim/21_project/21_project_Dresden_2009.pdf

STATISTICS EDUCATION RESEARCH JOURNAL REFEREES DECEMBER 2006-NOVEMBER 2007

Acknowledgement: We are grateful to the following who acted as referees for the *Statistics Education Research Journal* in the past year.

John Baker, USA
Mustafa Baloglu, Turkey
Dani Ben-Zvi, Israel
Monique Bijker, The Netherlands
Carol Joyce Blumberg, USA
Manfred Borovcnik, Austria
Lea Bregar, Slovenia
Bob delMas, USA
Scott Evans, USA
Larry Feldman, USA
Jenny Freeman, UK
Joan Garfield, USA
Randall Groth, USA
John Harraway, New Zealand
Tim Jacobbe, USA
Peter Johnston-Wilder, UK
Maria Kateri, Greece
Sibel Kazak, USA
Carolyn Keeler, USA
Dave Krantz, USA
David Lane, USA
Carl Lee, USA
Hollylynne Stohl Lee, USA

Nancy Leech, USA
Jiajuan Liang, USA
Helen MacGillivray, Australia
Denise Mewborn, USA
Nyaradzo Mvududu, USA
Ann O’Connell, USA
Irena Ograjenšek, Slovenia
Tony Onwuegbuzie, USA
Lionel Pereira-Mendoza, Canada
Susan Peters, USA
Peter Petocz, Australia
Dave Pratt, UK
Chris Reading, Australia
Anna Reid, Australia
Ernesto Sanchez, Mexico
Candace Schau, USA
Richard Scheaffer, USA
Mike Shaughnessy, USA
Eric Sowey, Australia
Kostas Triantafyllopoulos, UK
Angustias Vallecillos, Spain
Barbara Ward, USA
Jane Watson, Australia