Statistics Education Research Journal

Volume 5 Number 1 May 2006

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>. These files should be produced using the Template available online. Full details regarding submission are given 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

ISSN: 1570-1824

SERJ EDITORIAL BOARD

TABLE OF CONTENTS

EDITORIAL

The new year has brought transitions to SERJ. Flavia Jolliffe and Chris Reading have ended their terms as Co-editor and Assistant Editor, respectively. We are pleased to report that both have agreed to continue to share their expertise with the statistics education research community as SERJ associate editors. Thanks to both of them for their tireless work to promote SERJ and ensure the highest standards of quality in peer review and publication. Flavia deserves special thanks for shepherding the articles that appear in this issue of SERJ through the review process.

We welcome Beth Chance from California Polytechnic University, San Luis Obispo as the new SERJ Assistant Editor. Beth will work with the editors and authors to prepare the SERJ articles and announcements for publication. We are very grateful that Beth is willing to share her time and technical expertise.

We also welcome Peter Petocz as a new SERJ Associate Editor. Peter is currently a faculty member at Macquarie University in Sydney, Australia. We are eager to provide opportunities for Peter to share his extensive knowledge and experience within the SERJ manuscript review process and in guiding the policies and procedures of SERJ through service on the editorial board.

SERJ has already received 15 original manuscripts in the first five months of 2006. This is an increase over past years, and represents the growing awareness of and respect for peer reviewed statistics education research. We will continue to work to foster growth both in the number and quality of SERJ submissions.

We hope that the upcoming International Conference on Teaching Statistics (ICOTS-7) to be held in Salvador (Bahia), July 2-7, 2006, will provide researchers with not only the opportunity to present their work and publish in the conference proceedings, but to also submit a more elaborate report to SERJ. We encourage interested authors to consult our author guidelines, and also attend to the advice there regarding needed differences between Proceedings papers and more elaborate papers submitted for SERJ review, in order to avoid duplicate publishing.

Several SERJ activities are being planned for ICOTS-7. A lunchtime workshop on July 7th is designed for prospective authors; it focuses on writing and publishing research papers in peer-reviewed journals such as SERJ and on ways to avoid common problems found in manuscripts being submitted for review. Another lunchtime workshop on July 3rd is designed for current and future referees of papers submitted for review; it will examine the role of a referee in the scientific review process, and make suggestions for improving referees' work in light of what is sought in high-quality papers and the 'critical yet supportive' spirit expected in a referee report. If you intend to attend ICOTS and consider publishing research papers or helping as a journal referee, make sure to include these events in your planning.

The current issue of SERJ represents the first issue in the fifth year of the journal's existence. We feature three articles on a variety of topics. The article by Boyd Enders and Diener-West illustrates how a randomized trial can be used in education research, specifically to demonstrate that active learning can enhance student learning in an introductory biostatistics course. Lecoutre, Rovira, Lecoutre, and Poitevineau present empirical results to compare the understanding of randomness between teenagers, psychology researchers, and mathematics researchers. They found that in all three groups there were differences in understanding depending on whether the example was "real,"

Statistics Education Research Journal, 5(1), 2-3, http://www.stat.auckland.ac.nz/serj

that is, associated with a realistic context, or "stochastic," which means that the context involved traditional devices such as coins and dice. With an article about pre-service elementary school teachers' conceptions about variability, Canada provides a natural follow-up to the Fall 2004 theme issue of SERJ on "Research on Reasoning about Variability." The population of pre-service teachers provides fertile ground for further statistics education research. Alldredge and Brown use a non-randomized experiment to explore the effects of technology and student gender on pre-course beliefs and student learning. The surprising result is that the association between beliefs and learning evolved throughout the course. We hope that these articles will inspire others to ask "I wonder …" questions and initiate their own research agendas.

In November 2006 we anticipate the publication of a Special Issue on Reasoning about Distributions, with guest editors Chris Reading and Maxine Pfannkuch.

Thanks to our readers, authors, Editorial Board members, reviewers, and IASE and ISI leadership for your enthusiastic support for SERJ!

IDDO GAL AND TOM SHORT

MESSAGE FROM FLAVIA JOLLIFFE

As this is the first issue of SERJ where I have not acted as co-editor, I should like to take the opportunity of reminding readers of a few key developments, and of thanking past and present members of the editorial board for their support and for their tolerance of my nagging. I should also like to thank the many colleagues from around the world who have submitted papers or acted as referees. Without all these people there would be no journal.

I was a founding co-editor with Carmen Batanero and we and five others started planning what was then a new electronic journal, replacing the Statistics Education Research Newsletter from the Summer of 2001. The first issue was published in May 2002 and the editorial board was expanded later that year. Iddo Gal replaced Carmen as co-editor in December 2003. It has been a privilege for me to work with both of them, and also with Chris Reading whose period as assistant editor responsible for the production of issues and the web site has coincided with my time as co-editor.

There are many decisions to make when a new journal is started and SERJ has come a long way in the last five years. It is, however, still evolving. It is my hope that before too long we shall have so many publishable papers submitted to us that we shall grow to three issues a year.

FLAVIA JOLLIFFE

NEW EDITORIAL BOARD MEMBERS

SERJ is pleased to announce the appointment of **Beth Chance** as SERJ's new Assistant Editor. Beth is Associate Professor of Statistics at California Polytechnic State University in San Luis Obispo, California, USA. Her research interests include effective integration of technology into statistics courses and use of authentic assessment in introductory statistics. In addition to her research publications, she is co-author of the *Workshop Statistics* series (Key College Publishing), an innovative text that fully incorporates active learning and student discovery in introductory statistics, and of *Investigating Statistical Concepts, Applications, and Methods* (Duxbury Press). She served as an associate editor for textbook reviews for the *Journal of the American Statistical Association* (1998-2000) and as co-editor of *STATS: The Magazine for Students of Statistics* (2002-2004). She was the inaugural recipient of the American Statistical Association's Waller Education Award for innovation and excellence in teaching introductory statistics and was named a Fellow of the American Statistical Association in 2005. Beth has already been in communication with Chris Reading, who volunteered to offer further support during the transition process. Please join us in welcoming Beth to the SERJ Board.

SERJ welcomes **Peter Petocz** as a new SERJ Associate Editor. Peter is Associate Professor in the Department of Statistics at Macquarie University in Sydney, Australia, and Associate Dean for Teaching and Learning in the Division of Economic and Financial Studies. He is the author of several textbooks and video-based learning packages in statistics and mathematics, and has written about the development and use of appropriate learning materials. He has a research interest in statistics pedagogy in general, and more particularly in students' and teachers' conceptions of statistics and learning, and higher-level graduate "dispositions" such as creativity, ethics, sustainability and cross-cultural sensitivity. He has published widely on these topics, mostly with his research partner Anna Reid. Peter has a PhD in the area of stochastic processes, and is involved with several applied statistical studies mainly in the areas of orthodontics, nutrition and diabetes.

METHODS OF LEARNING IN STATISTICAL EDUCATION: A RANDOMIZED TRIAL OF PUBLIC HEALTH GRADUATE STUDENTS

FELICITY BOYD ENDERS, PHD
Mayo Clinic, Division of Biostatistics
enders.felicity@mayo.edu

MARIE DIENER-WEST, PHD
Johns Hopkins University, Department of Biostatistics
mdiener@jhsph.edu

ABSTRACT

A randomized trial of 265 consenting students was conducted within an introductory biostatistics course: 69 received eight small group cooperative learning sessions; 97 accessed internet learning sessions; 96 received no intervention. Effect on examination score (95% CI) was assessed by intent-to-treat analysis and by incorporating reported participation. No difference was found by intent-to-treat analysis. After incorporating reported participation, adjusted average improvement was 1.7 points (-1.8, 5.2) per cooperative session and 2.1 points (-1.4, 5.5) per internet session after one examination. After four examinations, adjusted average improvement for four study sessions was 5.3 points (0.4, 10.3) per examination for cooperative learning and 8.1 points (3.0, 13.2) for internet learning. Consistent participation in active learning may improve understanding beyond the traditional classroom.

Keywords: *Statistics education research; Active learning; Cooperative learning; Internet learning; Randomized trial*

1. INTRODUCTION

The discipline of statistics provides critical quantitative tools for public health researchers and practitioners. Students pursuing graduate degrees in public health must become familiar with key concepts in statistical reasoning and knowledge of the appropriate use and interpretation of classical biostatistical methods such as estimation, hypothesis testing, and multivariable analysis. In particular, the widespread availability and accessibility of statistical computing has increased the potential for public health professionals to confront statistical analyses in published reports, perform their own data analyses, or collaborate with research teams.

Because of their quantitative nature, courses covering statistical concepts and methods may be challenging for students from other fields of study. A variety of reasons have been proposed to explain why students might have difficulty in developing introductory statistical skills and competencies. Such students frequently harbor long-held anxiety regarding mathematical courses and traditional didactic teaching methods may not allow them to sufficiently overcome such fears (Bradstreet, 1996). In addition to these barriers, students are often enrolled in multiple courses or concurrently employed,

leading to a stressful background environment (Simpson, 1995). Finally, courses in introductory statistics draw such a variety of students from diverse backgrounds and with different prior knowledge and innate skills that it can be exceedingly challenging for instructors to simultaneously tailor didactic course material to meet all student needs (Simpson, 1995).

Recent advances in educational psychology and computer technology suggest possible ways to improve students' conceptual understanding of key statistical concepts. New instructional methods may enhance statistical education and students' learning of statistical concepts. One way to tailor statistical education is to include active learning methodology. "Active learning" refers to engaging a student in an activity, as contrasted with a lecture format or textbook which solely provides the student with information. A review of the literature in statistical education reveals that students may learn more readily when material is presented through student interaction or activities, as compared to the traditional passive lecturing style (Bradstreet, 1996; Garfield, 1995a; Garfield, 1995b; Lovett & Greenhouse, 2000; Moore, 1997). Ideally, this direct interplay forces students to overturn misconceptions, fears, or learning difficulties that hamper their ability to develop correct statistical intuition (Garfield, 1995a; Garfield, 1995b; Lovett & Greenhouse, 2000). Including such methodologies in the learning process might help improve students' understanding of statistical concepts. By establishing a "hands-on" environment, active learning may help alleviate difficulties heightened by anxiety related to mathematical concepts.

Active learning can be facilitated in a number of ways. "Cooperative learning" is accomplished when students work together on a structured activity in small groups to gain conceptual understanding (Garfield, 1993). This can be accomplished during, after, or instead of a traditional lecture. One method is to reinforce concepts and techniques introduced in a didactic lecture by subsequent small group activities facilitated by a teaching assistant. By working together, students not only engage in active learning, but derive benefits from their combined knowledge base.

Although the majority of previous attempts to implement active learning within statistical classrooms have used a cooperative learning approach (Gnanadesikan, Scheaffer, Watkins, & Witmer, 1997; Kvam, 2000; Magel, 1998), this might be difficult to accomplish with a large class size. Magel (1998) used cooperative learning in a class of 195 students and found it required significant advance preparation to break students into the small groups required and still have a single instructor serve as a facilitator for all the groups. Creating an interface with active learning using currently available internet technology provides an alternative approach for improving student understanding in large classes with a didactic course format. JAVA applets (mini-applications) provide a venue for students to independently examine statistical phenomena within a controlled internet-based environment. The interactive nature of the applets allows active learning to take place on the computer, i.e., "internet learning." Internet learning is distinct from "hybrid learning" (Utts, Sommer, Acredolo, Maher, & Matthews, 2003; Ward, 2004). In a hybrid course, the bulk of the course is online, and in person contact with students is limited, often to approximately an hour per week. By contrast, internet learning acts as an online component added to a traditional didactic course.

Previous studies have described the use of cooperative learning (Gnanadesikan et al., 1997; Kvam, 2000; Magel, 1998; Shaughnessy, 1977), but very few studies have compared cooperative learning or technologically-enhanced learning with the more traditional didactic or lecture-based style. This research study focuses on the implementation and evaluation of the addition of innovative instructional methods to an existing didactic course sequence in introductory biostatistics for non-statisticians. The

present study was designed to evaluate cooperative learning and internet learning within a randomized setting, and to compare the relative merits of cooperative and internet learning to each other and to a control group.

2. METHODS

2.1. STUDY DESIGN

This study was conducted from September through December 2001 (16 weeks) in the context of an introductory biostatistics course that was a requirement for students in most Masters and Doctoral degree programs at a school of public health. Standard course instruction included 3 hours of lecture and one 2-hour laboratory each week. The laboratory consisted of a structured review of examples pertaining to lecture material but in a smaller group setting that permitted more discussion. The first half of the course reviewed introductory concepts such as graphing, summary statistics, exploratory data analysis, probability concepts and distributions, and estimation and hypothesis testing. The second half of the course covered inference for one or two groups, analysis of variance, and simple linear regression. Learning materials consisted of lectures, accompanying lecture notes, laboratory exercises, problem sets, online self-evaluation problems, Stata™ (The Stata Corporation, 2001) computing notes, quizzes, and examinations.

The study design was a randomization among consenting students to one of three groups: cooperative learning (in person), internet-based learning (online), and control (see Figure 1 for a schema of the study design and participation). During the first week of classes, students were offered the opportunity to participate in the study and asked to complete an online pre-study survey of their mathematical and statistical skill and aptitude as well as demographic characteristics. All students were eligible, but were enrolled in the study only after providing written informed consent. In order to ensure representation of all degree programs, the randomization was stratified by degree program (Doctoral, MPH, other Masters degree, or Other). After randomization, the intervention phase was initiated. Each of the intervention sessions began after the introduction of the relevant concepts in lecture, and followed the same basic framework. Students in the cooperative learning group attended a one hour bi-monthly small group active learning session facilitated by a single experienced teaching assistant who did not participate in any other course-related activities. Many of the active learning sessions were motivated by projects described in *Activity-Based Statistics* by Scheaffer, Gnanadesikan, Watkins, and Witmer (1996).

At the same time, students in the internet learning group individually accessed a specially designed internet learning website and completed an internet-based activity typically focused on statistical concepts illustrated by interactive JAVA applets. The website was comprised of applets publicly available on the internet that were designed to help students learn particular statistical concepts. For each session, links to these applets were embedded in a single computer screen providing short instructions and questions for the students. The applets and their instructions remained available to students throughout the study.

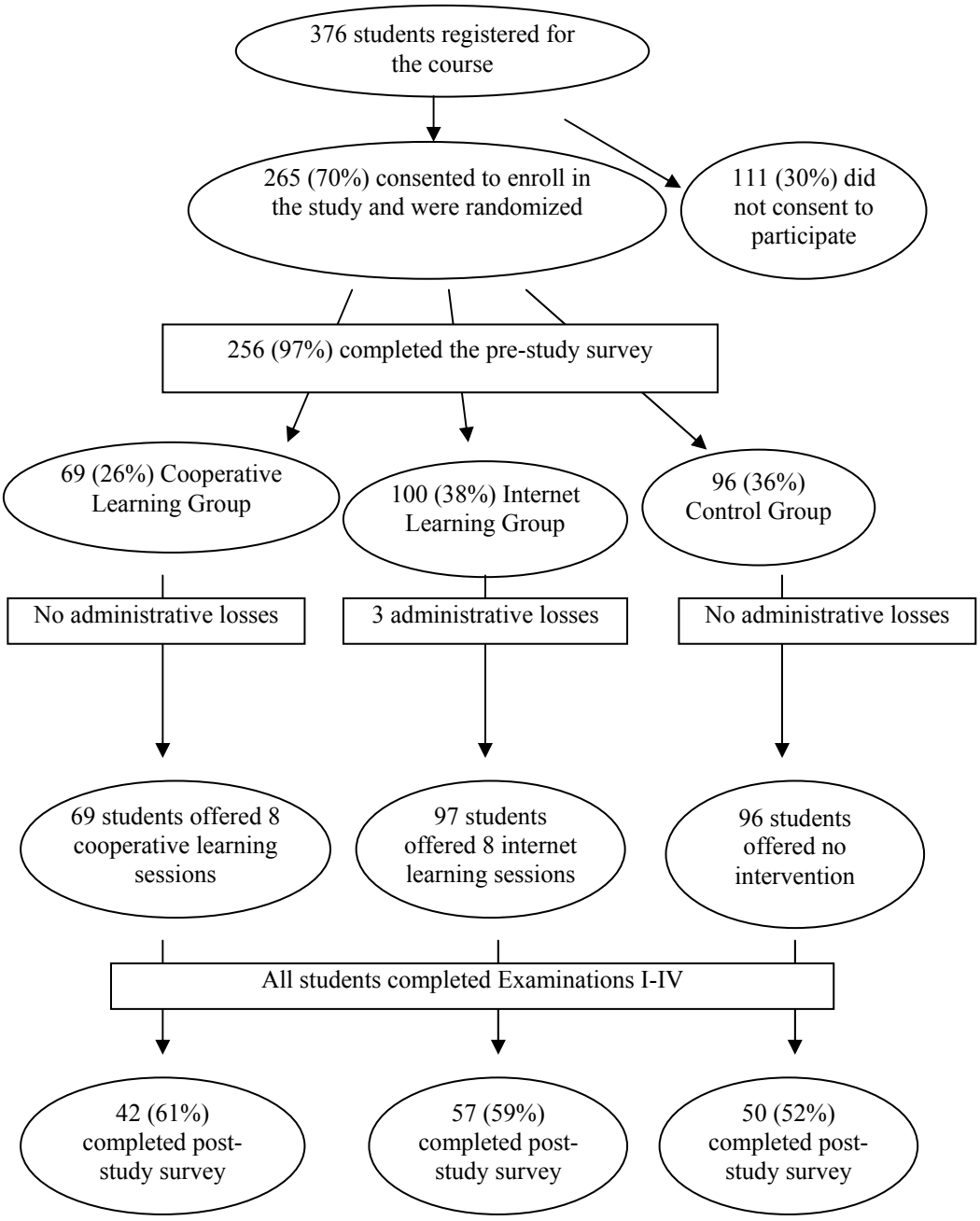

Figure 1. Study design and participation

The intervention sessions covered eight topics deemed integral to the understanding of course material: 1) conditional probability in a 2×2 table; 2) the Binomial and Poisson distributions; 3) the sampling distribution of the sample mean; 4) hypothesis testing; 5) confidence intervals; 6) the X^2 distribution; 7) analysis of variance; and 8) simple linear regression. Assessments were based on student performance as measured by four course examination scores. The first course examination was administered after the second study session; the second course examination was administered after the fourth study session; the third course examination was administered after the sixth study session; and the fourth and final course examination was administered after the eighth study session. Each

examination focused on material since the prior examination and included 20 five point questions so that possible scores ranged from 0 to 100 points.

2.2. STATISTICAL ANALYSIS

The primary goal of the analysis was to investigate possible associations between intervention and student performance in the course as measured by course examination scores. Three separate linear modeling approaches were used to compare student performance by study group (McCullagh & Nelder, 1989). In the first two approaches, the four examination scores for each student (0 to 100 points) were used as a set of four longitudinal outcomes with an exchangeable covariance structure; in the third approach, the outcome was the cumulative examination score (0 to 400 points). See Appendix A for equations used in each of the three approaches.

Intent-to-treat models: The first approach utilized the intent-to-treat principle; in Model 1, examination score was regressed on assigned study group. A random effect at the student level was employed for the repeated measures structure resulting from the use of the four examination scores for each student. Three indicator variables were included to adjust for variability in scores across the four course examinations (the fourth examination served as the reference).

Individual reported participation models: The second approach incorporated students' reported participation in the sessions. In Model 2a, participation in the most recent study session in either the cooperative learning group or the internet learning group or participation in neither session was used to predict the subsequent examination score. Model 2b used participation in both of the two most recent sessions. Similarly, Models 2c and 2d incorporated participation in the three most recent sessions (if available), or the four most recent sessions (if available), respectively. Since the intervention participation effects could vary by examination, two-way interaction terms between intervention participation and examination were added to the models shown in Appendix A. A random effect at the student level was employed for the repeated measures structure. Three indicator variables were included to adjust for variability across the four examinations.

Cumulative reported participation models: The third approach accounted for the total number of study sessions attended by each student in the intervention groups. Students in the control group were excluded from Model 3. Since cumulative participation in study sessions was not complete until the end of the study, the outcome for this approach was the sum of the four examination scores (the cumulative examination score). This approach estimated the effect of intervention on cumulative examination score after adjusting for the number of study sessions in which the student reported participation. Since only one observation per student was required for this analysis, no repeated measures structure was necessary.

The session participation used in the second and third modeling approaches was based on self-report either at the time of completion of the self-evaluation problems after individual study sessions or during the post-study survey. Because self-report was not requested at the time of the first study session, the first session was not included as a separate time-point.

Each model was subsequently adjusted for baseline factors associated with performance which were identified from the pre-study survey (data not shown). Non-consenting students were not included in analyses of examination scores, according to the regulations of our investigational review board. However, completion of the pre-study survey was taken as tacit consent for that portion of the study among students who did not consent to join the whole study.

3. RESULTS

3.1. STUDY PARTICIPATION

A total of 376 students registered in the course; 265 (70%) of the students consented to enroll in the trial with 69, 100, and 96 randomized to the cooperative learning, internet learning, and control groups, respectively. Three students randomized to the internet learning group were excluded from the analysis due to early changes in student course plans, reducing the total number to 97.

The distributions of demographic and student characteristics for both randomized and non-enrolled students are shown in Table 1. As expected by randomization, all three groups were fairly comparable with respect to pre-study characteristics, with no statistically significant differences. In addition, few differences were found between students who consented to join the study and those who did not enroll but who still completed the pre-study survey. Approximately 49% of the non-enrolled students voluntarily completed the pre-study survey. The primary difference between these two groups was that non-enrolled students reported greater levels of concurrent employment.

Individual access to the study interventions was not tracked in either intervention group, although self-report of intervention participation was collected. In the cooperative learning group, the number of students present at each session was collected. In the internet learning group, the Superstats™ software was used to track overall access to the internet learning website over time (SiteCatalyst, 2002). Figure 2 compares the overall participation by session from these two methods. However, since the method of tracking participation differed by intervention group, comparisons in Figure 2 can only be made regarding overall patterns of participation, rather than the participation rate, because of differences in scale. Of the 69 students randomly assigned to the cooperative learning group, 45 (65%) attended the first session on September 13, 2001, two days after a national tragedy in the US.

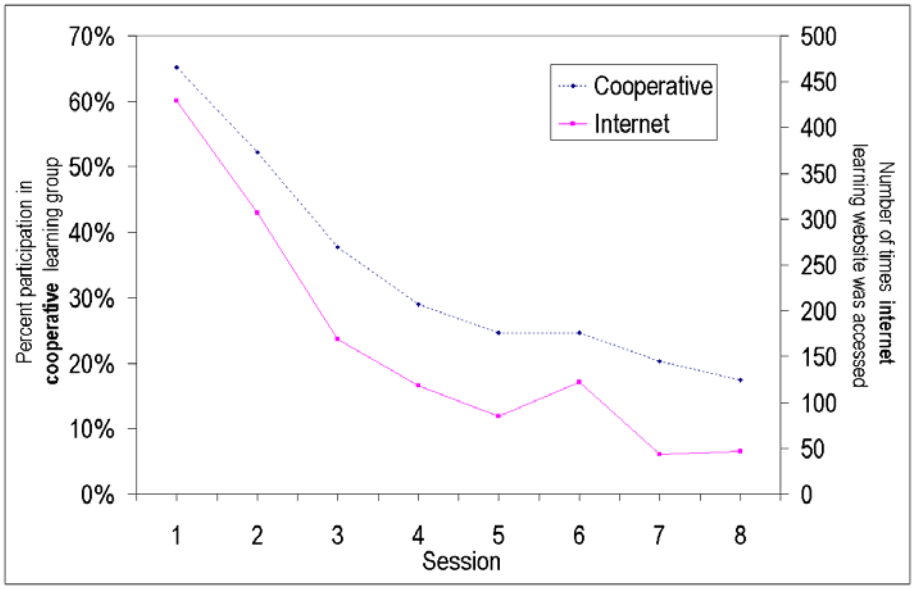

Figure 2. Participation in the cooperative learning and number of times the internet learning website was accessed, by study session

Table 1. Distributions of demographic and student characteristics for randomized and non-enrolled students

	Cooperative No. (%)	Internet No. (%)	Control No. (%)	p†	
Gender					
Male	20 (30.3)	27 (28.7)	30 (31.6)	0.91	19 (35.2)
Female	46 (69.7)	67 (71.3)	65 (68.4)		35 (64.8)
Age					
20-29	40 (58.0)	59 (60.8)	59 (61.5)	0.87	29 (53.7)
30-39	21 (30.4)	31 (32.0)	29 (30.2)		23 (42.6)
40-49	5 (7.3)	3 (3.1)	6 (6.3)		1 (1.9)
50+	3 (4.4)	4 (4.1)	2 (2.1)		1 (1.9)
Degree					
MPH	25 (37.9)	36 (38.3)	31 (32.6)	0.99	14 (25.9)
Other Masters	22(33.3)	32 (34.0)	36 (37.9)		19 (35.2)
Doctoral	12 (18.2)	17 (18.1)	17 (17.9)		12 (22.2)
Other	7 (10.6)	9 (9.6)	11 (11.6)		9 (16.7)
Credit Hours					
≤ 5	3 (4.4)	7 (7.2)	9 (9.4)	0.73	7 (13.0)
6-11	2 (2.9)	2 (2.1)	5 (5.2)		4 (7.4)
12-18	41 (59.4)	52 (53.6)	49 (51.0)		23 (42.6)
19+	23 (33.3)	36 (37.1)	33 (34.4)		20 (37.0)
English					
Native Language	42 (63.6)	52 (55.3)	60 (63.2)	0.45	32 (59.3)
Second Language	24 (36.4)	42 (44.7)	35 (36.8)		22 (40.7)
Employment					
10+ hours/week	24 (36.4)	30 (31.9)	41 (43.2)	0.28	44 (81.5)
<10 hours/week	42 (63.6)	64 (68.1)	54 (56.8)		10 (18.5)
	Mean (SD)	Mean (SD)	Mean (SD)	p‡	
Statistical Knowledge (Correct responses of 10)	4.28 (1.84)	4.32 (1.90)	3.71 (2.23)	0.078	4.44 (2.33)
Mathematical Skill (Correct responses of 5)	4.37 (1.22)	4.26 (1.23)	4.44 (1.02)	0.53	4.17 (1.19)
Desire for a Tutor (Likert scale: 0=definitely not needed to 4=definitely needed)	1.52 (0.96)	1.49 (0.89)	1.53 (1.11)	0.97	1.35 (1.07)
Total students	69	97	96		54

† Statistical significance for the difference between the randomized groups determined by Chi-square test.
‡ Statistical significance for the difference between the randomized groups determined by Analysis of Variance test.

3.2. STUDENT PREFORMANCE ON EXAMINATIONS

The overall mean cumulative examination score was 330.8 points (SD: 36.8 points). There was variability in mean score across the four course examinations. The overall mean (SD) scores were 89.0 (12.8) points; 81.3 (11.7) points; 82.8 (10.6) points; and 75.7 (14.1) points for the first through fourth examinations, respectively.

In a previous analysis variables from the pre-study survey were used to model cumulative examination score using forward stepwise regression incorporating two-way interaction terms. Younger age, greater mathematical aptitude (measured on a Likert scale from 0 to 5 based on a weighted scoring of the correct responses to questions 12 and 13 from Kemeny/Kurtz Math Series, 1992, p. 16) and statistical knowledge (measured on a 10 point scale adapted from Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000, and Wulff, Anderson, Brandenhoff, & Guttlet, 1987), working less than 10 hours per week, and student self-report of not needing a tutor (Likert scale of the reported need for a tutor; 0=definitely not, 1=probably not, 2=not sure, 3=probably, 4=definitely) were identified as pre-study factors associated with high performance. These five covariates were added in all subsequent models of intervention effect and performance.

Evaluating the Association of Intervention with Performance Based on Intent-to-Treat Models (1st Approach) No statistically significant differences in performance by study group were observed in the unadjusted intent-to-treat analyses. After adjusting for the five pre-study predictors of performance, estimated mean scores for students randomized to cooperative learning were 0.3 points below those of students randomized to control (95% CI: -3.4, 2.9); mean scores for students in the internet learning group were 0.01 points lower than students in the control group (95% CI: -2.8, 2.8).

Evaluating the Association of Intervention and Participation with Performance Based On Individual Reported Participation Models (2nd Approach) Table 2 shows results of the models of student performance on the four examination scores as a function of intervention and reported participation in sessions prior to the examinations. The results suggest increased performance in both intervention groups; however, statistically significant increases in performance were only observed at the time of the fourth examination. It should be noted that models for three or four consecutive study sessions could not be constructed for the first examination because only two study sessions had occurred by the time of that examination. All models in Table 2 included interaction terms of intervention effects and course examinations identified by Wald test results and were adjusted for the five pre-study predictors associated with performance.

Evaluating the Association of Intervention and Participation with Performance Based on Cumulative Reported Participation Models (3rd Approach) No statistically significant difference in performance (as measured by cumulative examination score) between the two intervention groups were observed after adjusting for the number of sessions the student reported attending (3rd Approach, see Table 3). However, performance increased with each additional study session in which the student participated. Each session was associated with a 2.1 point average increase (95% CI: 0.2, 3.9) in cumulative examination score in the adjusted model.

4. DISCUSSION

4.1. CONCLUSIONS

The goal of this study was to investigate whether the addition of active learning methods to a didactic introductory biostatistics course aided student understanding of key concepts, as measured by student performance on course examinations. The unadjusted intent-to-treat analysis revealed no statistically significant differences in performance across the three study groups (cooperative learning, internet learning, and control). This was likely attributable to low participation rates in the study interventions; by the third study session, 51% of the students in the two intervention groups had dropped out. From comments on the post-study survey, students in both intervention groups overwhelmingly cited lack of time as the predominant reason for nonparticipation. We also compared students who did not participate after the second study session with those who did complete later intervention sessions. The only difference found between participants and those who dropped out was that those completing later study sessions were enrolled in fewer credit hours.

In the presence of significant noncompliance, intent-to-treat analyses may not adequately reflect true differences between groups (Green, 2002). Accordingly, alternative analytic approaches were explored. The second modeling approach, using students' reported participation, suggested improved performance for participants as compared to nonparticipants and controls. The benefits of one study session were negligible. However, after four consecutive study sessions at the time of the fourth 100-point examination, cooperative learning participants scored an average of 5.3 points higher (95% CI: 0.4, 10.3), and internet learning participants scored an average of 8.1 points higher (95% CI: 3.0, 13.2), than nonparticipants or controls, after adjusting for the five pre-study factors associated with performance. The upper limit of the confidence interval reflects an improvement in understanding corresponding to perhaps two additional correct responses out of 20 examination questions. Under the 3rd modeling approach, each additional intervention session in which the student participated was associated with a 2.1 point increase in cumulative examination score (on a 400 point scale) (95% CI: 0.2, 3.9) in the adjusted model. When this effect is multiplied by the number of available intervention sessions, this increased performance may be substantial.

4.2. STRENGTHS AND LIMITATIONS

A limitation in the design of this study that could introduce bias was the requirement of extra work beyond the regular course material for the two intervention groups. One effect was decreased participation over time, which is associated with two potential biases; possibly students who continued to participate were more dedicated and thus more likely to work hard, or students who continued to participate did so because the intervention was more helpful to them than to those who dropped out. The effects of these potential biases may be most clearly observed in Table 2. By the time of the fourth examination, those who were still participating in the study interventions had likely participated in all four most recent study sessions; consequently, very little variation is observed in the increase in estimated performance from the models reflecting at least one study session versus the models reflecting at least four study sessions. Conversely, it is possible that students participating in the two intervention groups simply spent more time working with statistical concepts, and that additional time of any form would have led to the same improved performance.

Table 2. Linear models for students' subsequent examination scores by the number of prior study sessions attended (2nd Approach, interaction model)

	Intervention Group	Number of consecutive sessions attended	Change in Examination Score (95% CI)			
			Unadjusted Estimate		Adjusted Estimate*	
1st Examination	Cooperative Group vs. No Intervention	1 session	2.8	(-1.8, 7.3)	1.7	(-1.8, 5.2)
		2 sessions	4.8	(-0.3, 9.9)	2.6	(-1.6, 6.7)
	Internet Group vs. No Intervention	1 session	3.4	(-0.0, 6.9)	2.1	(-1.4, 5.5)
		2 sessions	3.1	(-0.4, 6.6)	1.7	(-1.7, 5.2)
2nd Examination	Cooperative Group vs. No Intervention	1 session	3.2	(-1.1, 7.4)	3.0	(-0.6, 6.6)
		2 sessions	3.4	(-1.2, 8.0)	3.1	(-0.8, 6.9)
		3 sessions	3.5	(-1.0, 8.0)	2.9	(-1.0, 6.8)
		4 sessions	4.1	(-0.4, 8.6)	2.9	(-1.2, 7.1)
	Internet Group vs. No Intervention	1 session	-0.7	(-4.0, 2.7)	-0.9	(-3.8, 2.0)
		2 sessions	-0.4	(-4.2, 3.3)	-0.9	(-4.3, 2.4)
		3 sessions	-0.2	(-4.0, 3.6)	-0.8	(-4.2, 2.6)
		4 sessions	-0.4	(-4.5, 3.7)	-0.9	(-4.7, 2.8)
3rd Examination	Cooperative Group vs. No Intervention	1 session	0.8	(-3.4, 5.0)	1.2	(-2.6, 5.0)
		2 sessions	1.3	(-3.2, 5.9)	1.7	(-2.4, 5.9)
		3 sessions	1.6	(-3.1, 6.2)	1.8	(-2.3, 5.9)
		4 sessions	1.6	(-3.0, 6.3)	1.7	(-2.4, 5.8)
	Internet Group vs. No Intervention	1 session	1.9	(-1.7, 5.4)	1.7	(-1.7, 5.2)
		2 sessions	2.3	(-1.4, 6.0)	2.1	(-1.5, 5.7)
		3 sessions	2.4	(-1.3, 6.3)	2.2	(-1.4, 5.8)
		4 sessions	2.6	(-1.3, 6.4)	2.1	(-1.6, 5.7)
4th Examination	Cooperative Group vs. No Intervention	1 session	3.6	(-2.7, 10.0)	4.5	(-0.4, 9.4)
		2 sessions	3.6	(-2.9, 10.0)	4.6	(-0.5, 9.6)
		3 sessions	4.2	(-2.3, 10.7)	*5.2*	*(0.1, 10.3)*
		4 sessions	4.3	(-2.2, 10.7)	*5.3*	*(0.4, 10.3)*
	Internet Group vs. No Intervention	1 session	*8.1*	*(2.9, 13.3)*	*7.1*	*(2.3, 11.9)*
		2 sessions	*8.8*	*(3.2, 14.4)*	*7.7*	*(2.6, 12.8)*
		3 sessions	*8.9*	*(3.3, 14.5)*	*7.9*	*(2.8, 12.9)*
		4 sessions	*9.0*	*(3.4, 14.6)*	*8.1*	*(3.0, 13.2)*

**Adjusted models include pre-study factors associated with performance.*

Table 3. Linear models for students' cumulative examination score by the number of prior study sessions attended (3rd Approach)

Comparison	Change in Cumulative Examination Score (95% CI)	
	Unadjusted Estimate	Adjusted Estimate*
Cooperative Group vs. Internet Group	0.4 (-11.4, 12.2)	4.7 (-5.3, 14.7)
Each additional study session	1.9 (-0.3, 4.0)	*2.1 (0.2, 3.9)*

**Adjusted models include pre-study factors associated with performance.*

Another potential bias is the Hawthorne effect. Individuals who are aware that they are being studied may behave differently than they otherwise would (Franke & Kaul, 1978). The average examination score was 1.5 points higher among those randomized to the control group (95% CI: -2.0, 5.0), than among those who did not consent to enroll in the study. However, it was not possible to adjust this difference for the other performance predictors, since not all non-enrolled students completed the pre-study survey.

The key strength of this study design was the randomization of students to different study groups. Utilizing a longitudinal framework allowed comparison of the effects of the interventions on performance over time. The large initial sample size of the trial provided the power necessary for detecting differences in performance by intervention and participation in our analyses.

The inclusion and comparison of two different types of active learning (cooperative and internet) was another key component of this study. With the advent of distance education, technologically enhanced learning, such as interactive online applets, affords a new way to offer active learning within a distance-friendly format. In the internet learning group, no supervision was required and yet improved performance was observed that was comparable to that of the cooperative learning group, with far less intensive investment of instructor time. The website for the internet learning group required little resources other than providing an introductory interface and framework since publicly available applets were used. Development of new applets would have initial costs but require little maintenance and resources over time.

4.3. IMPLICATIONS FOR FUTURE INSTRUCTION AND RESEARCH

Conducting a randomized trial within the framework of a large class such as this was extremely challenging. Despite the large number of students who initially chose to join the study, overall participation was low. Given students' hectic schedules and demands on their time, participation in any optional educational research project will be limited. Increased participation is needed in future investigations. One option is to incorporate intervention materials as a required course component. A comparative study could be made of consecutive offerings of a course in which the second offering introduces required new material (such as active learning strategies) but otherwise the course remains the same. This approach was used by Smith (1998) to evaluate small group

cooperative learning projects. Such a study design assumes no differences in student composition and requires the same assessment tools over time. However, the inclusion of such a comparison group is a critical part of evaluation of new statistical education methods.

Another consideration in the future evaluation of statistical education techniques is the specification of the amount of course content under evaluation. In this study, the intervention sessions covered a proportionately small amount of the course content and time relative to the other time requirements of the course. Consequently, only small changes in overall performance could be expected and their detection would require large sample sizes. Our choice of examination scores as the primary outcome variable resulted in high variability. Although we intended to utilize specific self-evaluation problems to evaluate the individual study sessions, these problems were not mandatory and thus were not completed by the majority of students. Future investigations warrant the incorporation of a required assessment tool that targets specific concepts emphasized through the intervention. One way to address this concern is the use of a hybrid course for the online portion of the intervention; however, ethical issues arise surrounding randomizing students to different types of courses.

These findings suggest that students may be aided by learning introductory biostatistics material via interactive activities especially if such activities are a required course component and offered throughout the term or semester. Our findings of an association of continued improvement in performance with completion of additional active learning sessions in the third approach is particularly encouraging. Cooperative learning activities and pertinent technological aids may both be helpful additions to on-site statistical education by either enhancing learning and/or reducing anxiety related to mathematical concepts. Future research and evaluation is needed to elucidate these relationships. In addition, research on online active learning methodologies is also required in the area of distance education. Continued development and evaluation of statistical teaching methodologies are critical and timely. Increasing numbers of public health professionals are seeking skills in quantitative methods and are faced with the challenge of mastering knowledge of appropriate statistical techniques and applications. The widespread availability of computer technology, both within and outside the classroom, provides an unparalleled environment for innovation in statistical education to maximize the potential for learning.

REFERENCES

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician*, *50*(1), 69-78.

Franke, R. H., & Kaul, J. D. (1978). The Hawthorne experiments: First statistical interpretation. *American Sociological Review*, *43*(5), 623-643.

Garfield, J. (1993). Teaching statistics using small-group cooperative learning. *Journal of Statistics Education*, *1*(1). [Online: http://www.amstat.org/publications/jse]

Garfield, J. B. (1995a). Respondent: How should we be teaching statistics? *The American Statistician*, *49*(1), 18-20.

Garfield, J. (1995b). How students learn statistics. *International Statistical Review*, *63*(1), 25-34.

Gnanandesikan, M., Scheaffer, R. L., Watkins, A.E., & Witmer, J.A. (1997). An activity-based statistics course. *Journal of Statistics Education*, *5*(2). [Online: http://www.amstat.org/publications/jse]

Green, S. B. (2002). Design of randomized trials. *Epidemiologic Reviews*, *24*(1), 4-11.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*(5500), 2261-2262.

The Kemeny/Kurtz Math Series. (1992). *Algebra workbook*. USA: True Basic, Inc.

Kvam, P. H. (2000). The effect of active learning methods on student retention in engineering statistics. *The American Statistician*, *54*(2), 136-140.

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician*, *54*(3), 196-206.

Magel, R. C. (1998). Using cooperative learning in a large introductory statistics class. *Journal of Statistics Education*, *6*(3). [Online: http://www.amstat.org/publications/jse]

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123-165.

Scheaffer, R. L., Gnanadesikan, M., Watkins, A., & Witmer, J. A. (1996). *Activity-based statistics: Instructor resources*. New York: Springer-Verlag New York, Inc.

Shaughnessy, J. 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.

Simpson, J. M. (1995). Teaching statistics to non-specialists. *Statistics in Medicine*, *14*, 199-208.

SiteCatalyst (formerly *SuperStat*). (2002). Nashville, TN: Omniture, Inc. [Online: http://www.omniture.com/]

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education*, *6*(3). [Online: www.amstat.org/publications/jse/v6n3/smith.html]

The Stata Corporation. (2001). *Intercooled Stata version 7*, Houston Station, TX.

Utts, J., Sommer, B., Acredolo, C., Maher, M. W., & Matthews, H. R. (2003). A study comparing traditional and hybrid internet-based instruction in introductory statistics classes. *Journal of Statistics Education*, *11*(3). [Online: www.amstat.org/publications/jse/v11n3/utts.html]

Ward, B. (2004). The best of both worlds: A hybrid statistics course. *Journal of Statistics Education*, *12*(3). [Online: www.amstat.org/publications/jse/v12n3/ward.html]

Wulff, H. R., Anderson, B., Brandenhoff, P., & Guttlet, F. (1987). What do doctors know about statistics? *Statistics in Medicine*, *6*, 3-10.

FELICITY BOYD ENDERS, PHD, MPH
Division of Biostatistics
Division of Health Sciences Research
The Mayo Clinic
200 First Street, SW
Rochester, MN 55905, USA

APPENDIX A: VARIABLE DEFINITIONS AND MODELS USED FOR STATISTICAL ANALYSIS

Variables used in the models:

Y is the vector of examination scores for the four course examinations

(Exam I) through (Exam III) are indicator variables for the first three course examinations

(Coop) and (Internet) are indicator variables for randomization to the two study intervention groups

(Number) is the number of study sessions for which the student reported participation

Y_{cum} is the sum of all four course examination scores (cumulative examination score)

The following time-defined variables are each vectors of length seven, for which time is defined as t = 2, 3, 4, …8, representing study sessions two through eight.

$(Coop)_t$ and $(Internet)_t$ are the vectors of indicator variables for reported participation in the study intervention groups across study session two through eight (control group is the reference group)

$(Coop)_{t-1}$ and $(Internet)_{t-1}$ are the vectors of indicator variables for reported participation in the two study intervention groups at the time of the prior study session (control group is the reference group)

$(Coop)_{t-2}$ and $(Internet)_{t-2}$ are the vectors of indicator variables for reported participation in the two study intervention groups at the time of the second prior study session (with I=0 for $t \leq 2$) (control group is the reference group)

$(Coop)_{t-3}$ and $((Internet)_{t-3}$ are the vectors of indicator variables for reported participation in the two study intervention groups at the time of the third prior study session (with I=0 for $t \leq 3$) (control group is the reference group)

Y_t is the vector of examination scores for the first course examination after the current study session

$(Exam\ I)_t$ is the vector of indicator variables that the course examination following the current study session (at time t) was the first course examination [$I(Exam_t = Exam\ I)$]

$(Exam\ II)_t$ is the vector of indicator variables that the course examination following the current study session (at time t) was the second course examination [$I(Exam_t = Exam\ II)$]

$(Exam\ III)_t$ is the vector of indicator variables that the course examination following the current study session (at time t) was the third course examination [$I(Exam_t = Exam\ III)$]

1: Intent-to-treat model

$$Model1.\ E[Y] = \alpha_0 + \beta_1(Coop) + \beta_2(Internet) + \gamma_1(ExamI) + \gamma_2(ExamII) + \gamma_3(ExamIII) + \varepsilon_{intra-student} + \varepsilon_{inter-student}$$

2. Individual reported participation models

$$Model2a.\ E[Y_t] = \alpha_0 + \beta_1(Coop)_t + \beta_2(Internet)_t + \gamma_1(ExamI)_t + \gamma_2(ExamII)_t + \gamma_3(ExamIII)_t + \varepsilon_{intra-student} + \varepsilon_{inter-student}$$

$$\text{Model 2b. } E[Y_t] = \alpha_0 + \beta_1(\text{Coop})_t + \beta_2(\text{Internet})_t + \beta_3(\text{Coop})_{t-1} + \beta_4(\text{Internet})_{t-1} + \gamma_1(\text{Exam I})_t + \gamma_2(\text{Exam II})_t + \gamma_3(\text{Exam III})_t + \varepsilon_{\text{intra-student}} + \varepsilon_{\text{inter-student}}$$

$$\text{Model2c. } E[Y_t] = \alpha_0 + \beta_1(\text{Coop})_t + \beta_2(\text{Internet})_t + \beta_3(\text{Coop})_{t-1} + \beta_4(\text{Internet})_{t-1} + \beta_5(\text{Coop})_{t-2} + \beta_6(\text{Internet})_{t-2} + \gamma_1(\text{ExamI})_t + \gamma_2(\text{ExamII})_t + \gamma_3(\text{ExamIII})_t + \varepsilon_{\text{intra-student}} + \varepsilon_{\text{inter-student}}$$

$$\text{Model 2d. } E[Y_t] = \alpha_0 + \beta_1(\text{Coop})_t + \beta_2(\text{Internet})_t + \beta_3(\text{Coop})_{t-1} + \beta_4(\text{Internet})_{t-1} + \beta_5(\text{Coop})_{t-2} + \beta_6(\text{Internet})_{t-2} + \beta_7(\text{Coop})_{t-3} + \beta_8(\text{Internet})_{t-3} + \gamma_1(\text{Exam I})_t + \gamma_2(\text{Exam II})_t + \gamma_3(\text{Exam III})_t + \varepsilon_{\text{intra-student}} + \varepsilon_{\text{inter-student}}$$

3. Cumulative reported participation models

$$E[Y_{\text{Cum}}] = \alpha_0 + \beta_1(\text{Coop}) + \beta_2(\text{Number}) + \varepsilon_{\text{inter-student}}$$

PEOPLE'S INTUITIONS ABOUT RANDOMNESS AND PROBABILITY: AN EMPIRICAL STUDY

MARIE-PAULE LECOUTRE
ERIS, Université de Rouen
marie-paule.lecoutre@univ-rouen.fr

KATIA ROVIRA
Laboratoire Psy.Co, Université de Rouen
katia.rovira@univ-rouen.fr

BRUNO LECOUTRE
ERIS, C.N.R.S. et Université de Rouen
bruno.lecoutre@univ-rouen.fr

JACQUES POITEVINEAU
ERIS, Université de Paris 6 et Ministère de la Culture
poitevin@ccr.jussieu.fr

ABSTRACT

What people mean by randomness should be taken into account when teaching statistical inference. This experiment explored subjective beliefs about randomness and probability through two successive tasks. Subjects were asked to categorize 16 familiar items: 8 real items from everyday life experiences, and 8 stochastic items involving a repeatable process. Three groups of subjects differing according to their background knowledge of probability theory were compared. An important finding is that the arguments used to judge if an event is random and those to judge if it is not random appear to be of different natures. While the concept of probability has been introduced to formalize randomness, a majority of individuals appeared to consider probability as a primary concept.

Keywords: *Statistics education research; Probability; Randomness; Bayesian Inference*

1. INTRODUCTION

In recent years Bayesian statistical practice has considerably evolved. Nowadays, the frequentist approach is increasingly challenged among scientists by the Bayesian proponents (see e.g., D'Agostini, 2000; Lecoutre, Lecoutre & Poitevineau, 2001; Jaynes, 2003). In applied statistics, "objective Bayesian techniques" (Berger, 2004) are now a promising alternative to the traditional frequentist statistical inference procedures (significance tests and confidence intervals). Subjective Bayesian analysis also has a role to play in scientific investigations (see e.g., Kadane, 1996). Formal applications of Bayesian probabilities are also more and more common in everyday life situations. Let us mention for instance the Bayesian spam-filtering techniques and the "probability of

precipitation" given by Canada's weather service (explicitly defined by that organization as a *subjective* estimate of rain or snow).

This change seriously questions the common choice in mathematics education to emphasize the teaching of the "frequentist" conception of probability and statistics and to virtually ignore the alternative Bayesian conception. Teaching the Bayesian approach appears nowadays both desirable and feasible (Berry, 1997; Lecoutre, Lecoutre, Grouin, 2001; Albert, 2002; Lecoutre, 2006), as previous objections (e.g., Moore, 1997) are less and less defendable. This should invite us not to radicalize the opposition between the Bayesian and frequentist inferences but rather to consider their interplay (see Bayarri & Berger, 2004). However, this requires a change of emphasis in the role of probability and randomness.

A recent empirical study indicates that students in introductory statistics class are generally confused about the different notions of probability (Albert, 2003). Clearly, continuing to teach *only* the frequentist conception cannot reduce the confusion. This implies to students either that there is only one "correct" conception of probability or that the frequentist and Bayesian conceptions are competitive, which should not be the case (Vranas, 2001). Moreover, an exclusive focus on frequentist notions may conflict with the students' intuitions and representations about probability (see e.g., Hawkins & Kapadia, 1984). In any case, as emphasized by Konold (1991, p. 144), "the teacher cannot, by decree, enforce a normative view."

We assume that variants of the concept of randomness are at the heart of probabilistic and statistical reasoning. In frequentist inference, only the data are random. As in the prototypical problems used in the traditional teaching of probability (flipping a coin, drawing a chip from a jar…), the frequentist conception involves a sequence of repeated trials or an ensemble of "identically" prepared systems. There is always a well-defined reference set of cases. So this seems to make probability an "objective" property of the data (of the coin, the chip…) existing in nature independently of us. In this sense, the frequentist approach emphasizes an "observable randomness" that can be "produced" (*simulated*). Unfortunately, *empirical* frequencies are seldom available for the assignment of probabilities in real problems. As a result, assigning a frequentist probability to a single case event is not often easy, since it requires *imagining* a reference set of events or a series of repeated experiments. Considerable teaching difficulties with the frequentist inference come from the fact that data are considered as random *even after observation*.

In the Bayesian approach, the parameters are also considered as random, while data after observation are fixed quantities. We need to use another conception of probability. The Bayesian probability is the *degree of belief* (or *confidence*) in the occurrence of an event or a measure of the degree of plausibility of a statement. It can serve to describe "objective knowledge," in particular based on symmetry arguments or on frequency data. It can also be used to express a *personal* description of a state of knowledge, eventually incorporating *subjective* opinions (Savage, 1954; de Finetti, 1974), a notion that the frequentist conception rejects as being problematic. With the Bayesian approach it is not *conceptually* problematic to assign a probability to a single case event. Moreover, the Bayesian definition fits the meaning of the term probability in everyday language, and so the Bayesian probability theory appears to be much more closely related to how people intuitively reason in the presence of uncertainty.

The calculus of probability has been introduced to formalize randomness. In the XIXth century, in accord with Laplace's determinist conception of the world, randomness is the word given to the ignorance of a person in a determined universe (Laplace, 1951). Nature is knowable and yields to mathematical rules. Randomness is either euphemism for ignorance, or the expression of the limits of human perception and knowledge; it is

randomness when unknown. An alternative conception of randomness, that has been expressed as a fundamental principle in quantum physics, is that physical reality is irreducibly random; it is *randomness per se*.

The frequentist view of probability is often perceived as related to determinism. So, in his preface to the re-edition of Laplace's book on probability, Thom (1986, p. 8) wrote "Laplace est ouvertement 'fréquentiste' (Laplace is openly 'frequentist'), *comme il se doit* de quelqu'un qui postule le déterminisme universel [as someone who postulates universal determinism *must be*)" [italics added]. However, things are not so simple and Laplace is also often associated with the development of Bayesian ideas. One must admit that the concept of randomness is ambiguous and complex (Kac, 1983) and gives rise to various interpretations.

Consequently, what people — students as well as instructors — mean by randomness should be taken into account when teaching these topics. Since the early 1950s, psychologists have carried out extensive research on people's ability to *produce* or *perceive* randomness. In the first case, subjects were required to simulate a series of outcomes of some typical random process such as tossing a coin (for a review, see especially Wagenaar, 1972). In the second case, subjects were asked to rate the degree of randomness of several sequences of stimuli. One of the main conclusions of all these studies is that humans are not good at either *producing* or *perceiving* randomness (Falk & Konold, 1997; Nickerson, 2002). However, all these studies strongly involved a frequentist conception of probability. This is also the case for the numerous studies using simulations of sampling distributions in order to improve students' statistical thinking processes.

So much research remains to be done to inform the teaching of Bayesian statistics. Konold et al. (1991) postulated that whether what students think is random, or not random, had a role in understanding probability distributions. In this perspective, they carried out an exploratory study on people's subjective criteria of randomness. Twenty psychology students and five mathematicians were asked to categorize familiar items as either "random" or "not random." The authors distinguished two types of items. "Stochastic" items either involve a repeatable process (e.g., rolling a die) or consist of outcomes produced *via* a mechanism associated with chance (e.g., drawing from a set of objects); by contrast "real" items consist of outcomes defined from everyday life experiences (e.g., the germination of a planted seed). The study found the following results: (1) A higher percentage of stochastic rather than real items was classified as random by both students and mathematicians; (2) The subjects' justifications showed a great diversity of conceptions; (3) Some mathematicians expressed their difficulty with having to dichotomize the items because they tended to view randomness as an entity that is present in degrees.

We assume that the spontaneous criteria for assessing randomness are linked to the theoretical definitions of probability. So, a stochastic item implicitly involves a unique (well-defined) reference set of cases, and consequently can be assigned to either a frequentist or Bayesian probability about which it can be expected that different individuals agree. On the contrary, a real item describes a single case event for which a well-defined reference set does not exist a priori. Consequently, an individual who doesn't accept Bayesian probabilities may consider that it is impossible to assign a probability to a real item, either because he/she has no reference set (a frequentist probability doesn't exist) or because he/she considers that any reference set should depend of his/her *personal* experience (an objective probability cannot be calculated). Consequently, greater variability can be expected for the real items. The psychological effect of considering a single case rather than a set of cases was termed the "power of the

particular" by Kahneman (quoted in Griffin & Buelher, 1999). It seems to encompass at least two separate phenomena: (1) More empathy and other emotional reactions are aroused with the single case, because it is easier to imagine and identify with; (2) It also invites analysis by reasoning processes that are case-specific and deterministic, rather than statistical.

With the purpose of further investigations, we devised a two-phase experiment in which the second phase was similar to the Konold et al. (1991) procedure, hence a "constrained categorization": here, *randomness was an imposed criterion*. It was preceded by a "free categorization" task: here there were *no imposed criteria*. We assumed that this task permitted us to gather answers as spontaneous as possible which should partly reflect the subject's representations and beliefs about randomness.

We compared three groups varying in expertise in probability: lower secondary school pupils, psychology researchers, and mathematical researchers. Our main objective was to provide evidence of some internal coherence in probability judgments.

2. METHODS

2.1. SUBJECTS

Three groups of 20 subjects participated in the experiment.

(1) COL group: 20 pupils of the third class of a "collège" in Rouen (in France this corresponds to the last class of the lower secondary school) were chosen at random. They were of both sexes and aged 14-16 years old. They had not had a course in either statistics or in probability.

(2) PSY group: 20 psychology researchers from universities in Rouen and Paris, all with a PhD. They were recruited if they had some training in probability and applied statistics and had some practical experience processing experimental data.

(3) MAT group: 20 mathematics researchers from the university of Rouen, all with a PhD. They were recruited if they had training in probability theory and in mathematical statistics and had experience in teaching probability or statistics.

2.2. MATERIALS

The 16 items, reported in Table 1, were presented on individual cards. They are a priori categorized into four classes. Eight real items are events from everyday life experiences. In 4 items the subject is implied in the formulation by the use of the personal pronoun "you," while this is not the case in the other 4 items. Eight stochastic items either involve a repeatable process or consist of events produced via a mechanism which is associated with chance. Four items involve two equally likely, symmetric outcomes, while the 4 other items involve asymmetric outcomes.

2.3. PROCEDURE

The subjects carried out the task individually. They were told that they would be taking part in an experiment aimed at assessing their spontaneous judgments on various familiar situations. The item cards were randomly mixed and simultaneously visible. First, the subjects were asked to "put together the cards which go together," and thus to make piles. They were told that they could make as many piles as they wished and could take as much time as they wanted (*free categorization*). After they completed this task, they answered the following question: "Why did you put those cards together, and those

ones, and so on?" Then the cards were mixed again and the subjects were asked to answer the following question: "For each card, do you think that there is randomness involved or not; explain why" (*constrained categorization*). The experiment lasted from 15 to 30 minutes.

Table 1. List of the 16 items

Real situations

With an implication of the subject in the formulation of the situation

A You meet a friend you have not seen for 10 years
B You win 10000F at the lottery
C You say the first thing that comes to your mind
D You will get the flu in the next month

Without any implication of the subject in the formulation of the situation

E A planted seed germinates or does not.
F The quotation of a stock at the Stock Exchange of Paris will go up more than 5% in the next three months
G It rained in Paris on March 15, 1936 or did not
H It will rain tomorrow in Paris

Stochastic situations

With symmetric outcomes

S An even number is obtained from a rolling of a die
T Heads is obtained from the toss of a fair coin
U Tails is obtained at the fifth flip of a fair coin that has landed with tails up on the previous four flips
V A white marble is drawn from a box that contains 10 black and 10 white marbles

With non-symmetric outcomes

W Two red chips are drawn from a box that contains 1 white chip and 2 red chips
X A pair of socks that match is obtained from a blind draw of two socks from a drawer in which there are two pairs of different socks
Y A lemon-flavoured sweet is drawn from a box that contains 20 orange-flavoured and 10 lemon-flavoured sweets
Z A white marble is drawn from a box that contains 10 black and 20 white marbles

3. RESULTS

3.1. FREE CATEGORIZATION

The categorizations were analyzed using the additive similarity trees (AST) model (Sattath & Tversky, 1977). This is a valuable alternative to multidimensional scaling (MDS), in which each object is represented by a point in a multidimensional (usually Euclidean) space. In AST the objects are represented by the external nodes of a tree. Roughly speaking, in the first step a topology is found such that a condition called the "four-point condition" (Buneman, 1974) is verified at best, in a certain sense. This condition, stronger than the triangle inequality, is characteristic of an additive tree. The

distances d between any four points x, y, z, t of the tree satisfy the inequality $d(x,y)+d(z,t) \leq max[\,d(x,z)+d(y,t),\, d(x,t)+d(y,z)\,]$. Then arc-lengths are scaled so that the length of the path joining two nodes has a close fit to the similarity between the corresponding objects. The four-point condition is weaker than the ultrametric (or strong triangle) inequality that must be satisfied by ultrametric trees associated with hierarchical clustering. Consequently, additive trees are more likely to provide a faithful representation of proximity data than ultrametric trees. Pruzansky, Tversky, and Carroll (1982) reanalyzed proximity data sets from various published studies and concluded that MDS was more appropriate when the hypothesized structure of the objects was perceptual and that AST was more appropriate when it was conceptual.

From the categorizations made by the subjects, an overall distance matrix was obtained; the distance between two items was the percentage of subjects who classified these two items into separate categories (thus the possible maximal distance between two items was 100). This matrix was input to the computer program ADDTREE (in the version by Barthélemy & Guénoche, 1991) to produce the additive similarity trees for the items.

The theoretical additive similarity tree associated with the a priori classification of the items is shown in Figure 1. The observed trees within each of the three groups of subjects are shown in Figure 2.

Figure 1. Theoretical additive similarity tree associated with the a priori *classification of the items into four classes.*

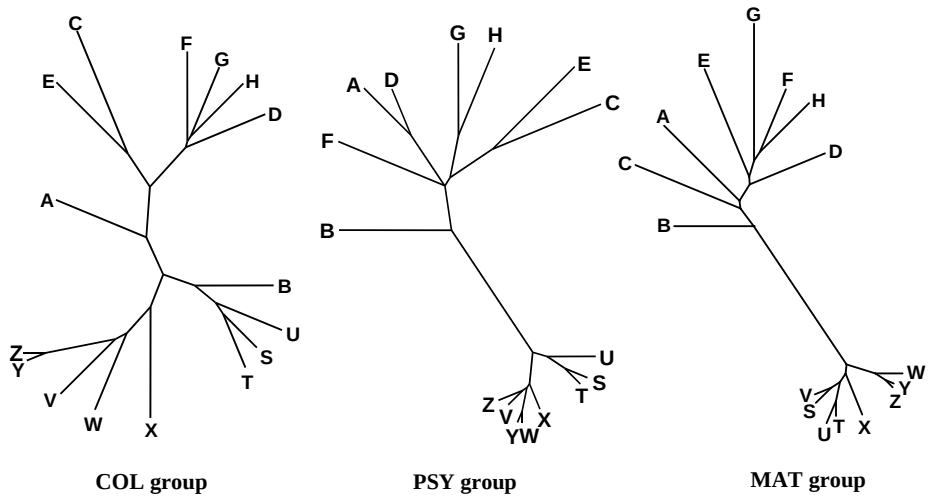

Figure 2. Free categorization: observed additive similarity trees within each of the three groups of subjects.

The tree-structures are quite similar for the three groups. The 16 items are first partitioned into two major clusters: real items versus stochastic items. The results are especially striking for the MAT group since all the real items (ABCEDFG) on the one

hand and all the stochastic items (STUVWXYZ) on the other hand are regrouped at two extremities of the tree. Furthermore in the MAT group, each of these two clusters is further partitioned into two finer clusters. The eight real items are divided according to the degree of implication of the subject: EFGH (without any implication) are separated from ABC (with implication), D (flu) being more distant. The eight stochastic items are divided according to the nature of their outcomes: STUV (with symmetric outcomes) are separated from WXYZ (with non-symmetric outcomes). For the two other groups of subjects, although the same first partition into two major clusters can be observed, the partition into finer clusters is less apparent, and the trees are more widespread, especially for the COL group. It can be noted that item B (lottery) is closer to the stochastic items than to the real items for the PSY group, and is clearly associated with the stochastic items for the COL group.

The *d* distance of Robinson and Foulds (1981) between the trees is reported in Table 2. This distance is purely topologic, that is to say that it only takes into account the structure of the trees, while ignoring the length of the paths. It is equal to the minimum number of elementary operations (fusion or division of nodes) necessary to transform one tree into another one, and for *k* items lies between 0 and 2*k*-6, thus here $0 \leq d \leq 26$.

Table 2. The d distance of Robinson and Foulds ($0 \leq d \leq 26$)

	MAT group	PSY group	COL group
PSY group	22		
COL group	20	12	
Theoretical tree	12	16	18

It is for the MAT group that the distance between the theoretical tree and the observed tree is the smallest. So the free categorizations of the MAT subjects are those which fit the best with the a priori theoretical classification. The trees of the two other groups are nearly equidistant. Furthermore, it can be noted that the maximal distances (nearly equal) are observed between the tree of the MAT group and the trees of the other two groups.

The justifications given by the subjects support further comments about the categorizations. There is large inter-individual variability, since there are almost as many different sets of categorizations and justifications as there are subjects. Nevertheless, a striking finding is that most subjects have explicitly used the notion of randomness in the free categorization task, although this term was never mentioned in the instructions. The main criterion used in all three groups involves the opposition between the items linked to a probability ("computable events") and those linked to everyday life experiences for which it is difficult if not impossible to calculate a probability. Other criteria are specific to each group. In the MAT group, we frequently observed an opposition between the events which are typical examples of standard models of randomness ("typical mathematical problems for our students") and the events in which there is randomness "when unknown" ("no available standard model"). Furthermore, within this group some categorizations are based either on the type of probability involved – conditional or elementary – or on the probability value (e.g., <1/2, 1/2, 1/3,…). These criteria can be viewed as variants of the aforementioned main criterion. In the two other groups there is greater inter-individual variability. In the PSY group, some subjects differentiated the events that are linked to nature or the environment (yielding to some meteorological or biological rules) from the "purely random" events. In the COL group, some categorizations are specifically based either on the opposition between *lucky* ("that's

chance") and *unlucky* ("that's fatality") events, or to the degree of implication of the subject.

3.2. CONSTRAINT CATEGORIZATION

A mathematician systematically answered "I don't know" to all the items and was eliminated from the study. Some subjects, rather than expressing a dichotomous attitude, rated graduated judgments such as "randomness is involved, but only a little." Consequently, the answers were a posteriori classified into three main categories: R (Random), L (a Little bit random) and N (Not random).

Trees The three additive trees for the items are reported in Figure 3. The tree-structures are quite similar for the three groups. On the whole the 16 items are partitioned into two major clusters: real items versus stochastic items. Nevertheless, there are two main exceptions concerning items A (friend) and B (lottery) which are separated from the other real items and are closer to the stochastic items.

A comparison of the trees obtained in the two phases, shows that making the notion of randomness explicit has three main effects: (1) For the stochastic items, the distinction between symmetric and non-symmetric outcomes is less apparent in the constrained categorization. So these two categories of items are perceived as quite similar when randomness is an explicit criterion of classification, while they can be perceived as different in the free categorization task; (2) In the constrained categorization the real items are much more dispersed, revealing more divergent conceptions when randomness is an explicit criterion; (3) The real items A (friend) and B (lottery) are relatively isolated and are much closer to the stochastic items in the constrained categorization than in the free one.

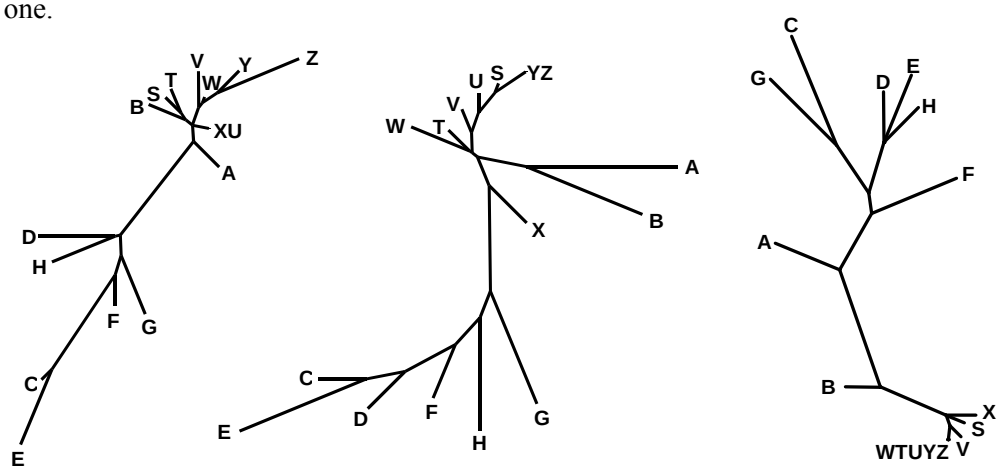

Figure 3. Constrained categorization: observed additive similarity trees within each of the three groups of subjects.

The justifications given by the subjects support further comments. For the stochastic items, the explicit reference to a random model in the constrained categorization is so salient that the distinction between symmetric and non-symmetric outcomes vanishes. For the two real items A and B in the free categorization, the presence/absence of implication is the salient property for most subjects in the three groups, which led them to classify

these two items with other real items. In the constrained categorization, it is the extreme unusualness of these two events that became the salient property for most subjects. Consequently these two items are separated from the other real items and are closer to the stochastic items which are viewed in majority as events in which randomness intervenes.

Inferential analyses The percentages of answers within the three groups are reported in Table 3. For the real items, there is no consensus except for a certain extent A (friend) and B (lottery), which are mainly categorized as random, especially by the COL subjects. For the other real items, there is no large systematic majority. However, they are mainly categorized as not random by PSY subjects (from 55% to 75%). Items C (first thing) and E (seed) are the most frequently categorized as not random (by respectively 69% and 67% of the subjects).

For each stochastic item, there is a large majority of COL and MAT subjects (from 85% to 100%) who categorize it either as random or in a few cases as a little bit random. Nevertheless, not one item is unanimously categorized as random by all these subjects. By contrast, each of these items is categorized as not random by a non negligible and approximately constant proportion of PSY subjects (from 30% to 40%).

Table 3. Proportions of answers for each of the 16 items within the three groups of subjects: COL (n=20)*, PSY* (n=20) *and MAT* (n=19)

		Real items								Stochastic items							
		With implication				Without implication				Symmetric outcomes				Asymmetric outcomes			
		A	B	C	D	E	F	G	H	S	T	U	V	W	X	Y	Z
R	COL	.90	.90	.25	.50	.05	.50	.45	.50	.90	.90	1	.85	.95	1	.85	.70
	PSY	.60	.65	.20	.25	.15	.25	.25	.25	.50	.50	.45	.45	.60	.45	.45	.45
	MAT	.53	.79	.16	.37	.26	.32	.16	.32	.89	.95	.95	89	.95	.89	.95	.95
	mean	.68	.78	.20	.37	.17	.36	.29	.36	.76	.78	.80	.73	.83	.78	.75	.70
L	COL	.05			.15			.05	.15							.10	.20
	PSY	.10	.20	.05	.15	.10	.10	.15	.20	.20	.10	.20	.15	.10	.15	.25	.25
	MAT	.26	.05	.16	.37	.42	.32	.21	.42				.05				
	mean	.14	.08	.12	.22	.17	.14	.14	.26	.07	.03	.07	.07	.03	.05	.12	.15
N	COL	.05	.10	.75	.35	.95	.50	.50	.35	.10	.10		.15	.05		.05	.10
	PSY	.30	.15	.75	.60	.75	.65	.60	.55	.30	.40	.35	.40	.30	.40	.30	.30
	MAT	.21	.16	.68	.26	.32	.37	.63	.26	.11	.05	.05	.05	.05	.11	.05	.05
	mean	.19	.14	.69	.40	.67	.51	.58	.39	.17	.18	.13	.20	.13	.17	.13	.15

R: Random; L: a Little bit random; N: Not random

In all the further analyses, the two categories “Random” and “a Little bit random” were grouped together and coded R (“Random”). A contrast analysis was performed in a three-groups ANOVA design with a four-level repeated factor, “Items,” corresponding to the four classes of items. The dependent variable was the proportions of items categorized as not random. Interval estimates are reported for each difference between proportions of interest. We used standard non-informative Bayesian procedures (see Lecoutre & Derzko, 2001) to assess either the largeness or the smallness of the population difference δ. Results are summarized in Table 4.

Table 4. Contrast analysis for the constrained categorization. The dependent variable is the proportions of items categorized as not random (N). *D is the observed difference and* δ *is the population difference.*

Contrast	*D*	*t* [df]	*p*	Bayesian conclusion
real versus stochastic	+.29	+7.06 [56]	<.0001	Pr(δ>+.24)=.90
real versus stochastic/COL	+.38	+6.38 [19]	<.0001	Pr(δ>+.30)=.90
real versus stochastic/PSY	+.20	+2.96 [19]	.0080	Pr(δ>+.11)=.90
real versus stochastic/MAT	+.30	+3.44 [18]	.0028	Pr(δ>+.18)=.90
implication versus no implication/real items	+.17	+4.78 [56]	<.0001	Pr(δ>+.13)=.90
symmetric versus asymmetric/stochastic items	+.03	+1.07 [56]	.2892	Pr(.01<δ<+.06)=.90
PSY *vs* COL+MAT	+.21	+2.96 [56]	.0044	Pr(δ>+.12)=.90

The observed difference between the unweighted average proportions of not random answers for the real (.45) and the stochastic items (.16) is D=+.29, significantly different from 0. The standard Bayesian analysis shows that there is a 90% probability that δ is larger than +.24; a notably large difference can be assessed. The same conclusions are found within each group of subjects (see Table 3).

Within the real items, the observed difference between the items without implication (.54) and those with implication (.36) is +.17 and a notably large difference can be assessed (Pr(δ>+.13)=.90). Within the stochastic items, the observed difference between the items with symmetric outcomes (.17) and those with asymmetric outcomes (.15) is +.03 (non significant) and a relatively small difference can be assessed (Pr(-.01<δ<+.06)=.90).

The observed difference between the average proportions of not random answers for the PSY group (.44) and the two other groups (.26 for the COL group and .21 for the MAT group) is +0.21. A notable difference can be assessed (Pr(δ>+.12)=.90) This difference is mainly attributable to the stochastic items that are more often categorized as not random within the PSY group than within the two other groups.

Individual patterns The individual patterns of the 16 answers given by each subject were further analyzed. Each pattern is a string of 16 Ns or Rs, ranked from items A to P. Taking into account the justifications given by the subjects, some general conceptions of randomness can be identified. Each identified conception defines a theoretical pattern. An example of a general conception is that randomness is involved whenever it is possible to calculate a probability; consequently, randomness is involved for stochastic items and is not involved for real items, hence the theoretical pattern NNNNNNNNRRRRRRRR.

We considered that an observed pattern was compatible with a theoretical pattern if at least 14 out of the 16 answers were the theoretical answers. This analysis shows that three general conceptions allow us to account for 75% (44/59) of the observed patterns.

(1) The majority conception is that randomness is involved whenever probabilistic reasoning is involved. Thus randomness is involved for *all items*. Thirty nine percent of the observed patterns (23/59) are compatible with a string of 16 Rs. This conception is more frequent in the MAT group (53%) than in the COL (35%) and PSY (30%) groups. Note that some subjects, especially in the MAT group, explicitly refer to two kinds of randomness: a "mathematical" randomness and a randomness "when unknown." Mathematical randomness would be linked to the events for which it is possible to compute a probability (typically the stochastic items); randomness when unknown would

be linked to all the events for which there is no possibility to *easily* compute a probability (typically the real items). For these subjects randomness is involved in *all* items, but stochastic and real items differ as to the *nature* of randomness.

(2) Another frequently encountered conception is that randomness is involved for the stochastic items – because it is possible to compute probabilities – and is not involved for the real items (except A, friend, and B, lottery, for which it would also be involved) because determinism plays a great part and causal factors can be identified. Thirty one percent (18/59) of the observed patterns are compatible with the corresponding theoretical pattern (RRNNNNNNRRRRRRRR) which is an approximate dichotomy between real and stochastic items. This is the majority conception in the COL group (40% compared to 26% in each of the two other groups).

(3) A conception encountered only in the PSY group is "randomness is never involved." Five percent (3/59) of the observed patterns are compatible with a string of 16 Ns. One of these three psychologists expressed a strong conviction that the world is entirely deterministic. The two other subjects stated that randomness is not involved whenever it is possible to compute a probability, or in their words "quantify."

The 25% remaining patterns involve partial conceptions which correspond to some specific views of randomness and apply only to some items. We will mention three of these conceptions. (1) A phenomenon is random only when all the outcomes have the same probability (cf. the "equally-likely" justification in Konold et al., 1991). Consequently randomness would be involved in at least the four stochastic items with symmetric outcomes; a typical justification is "it's pure random because we have 50/50 chances." This can be compared with the "equiprobability-bias" (Lecoutre, 1992) according to which random events are thought to be equiprobable "by nature" or with the "uniformity belief" (Falk, 1992) according to which people have a strong intuitive tendency to assume equal probabilities for the various available options. (2) A phenomenon is random when there is no prior knowledge about the outcome, and thus no possibility to predict nor to control the result (cf. the "causality" and "uncertainty" justifications in Konold et al., 1991). For instance E (seed germination) is "not random because one can control the soil, the wetness…", H (rain) is "not random, because there is a way of predicting the weather." By contrast, items involving a die or a coin are "random because one can't control or predict anything." These justifications can be compared with the theory of Piaget and Inhelder (1951) according to which the emergence of the idea of chance is attributed to children's realization of the impossibility of predicting oncoming events or of offering causal explanations. (3) Some justifications reflect a conception which connects the degree of intervention of randomness to the value of the probability. Randomness is said to be involved more as the probability decreases. For instance, W (chips) is "almost not random because the probability is relatively high;" by contrast B (lottery) "is really random, because the probability is very weak."

4. CONCLUSION

Our study has confirmed that individuals hold a wide range of meanings for the concept of randomness, since in the two categorization tasks taken altogether there were as many different classifications as there were subjects, however simple and familiar the 16 items may be. These findings are in accordance with the results of many studies which have been taken as evidence that the concept of randomness leads to a lot of different interpretations, even by many who use it extensively in their work (Nickerson, 2002). Nevertheless it was possible to distinguish some general conceptions of randomness, and so to provide evidence of some internal coherence in probability judgments. The 16 items

were partitioned into two main classes opposing the real and the stochastic items which were perceived as different. A large majority of individuals were in agreement for the stochastic items and categorized them as random because it is "easily" possible to compute a probability. In contrast individuals were divided for the real items; they categorized them either as random or not random with no large majority. Two main conceptions have been observed for the real items: Either randomness is involved because a probabilistic reasoning is involved, or randomness is not involved because determinism plays a great part or because causal factors can be identified. These findings are compatible with the "power of the particular" according to which the single cases "seem to invite analysis by reasoning processes that are case-specific and deterministic, rather than statistical."

Another important finding was the little effect of background knowledge of probability theory on one's views of randomness. In particular, the dichotomy of stochastic versus real items was observed within each of the three groups, including lower secondary school pupils without any background knowledge of probability theory. This is compatible with Konold's conclusion according to which students have strong intuitions about probability and randomness prior to instruction (Konold, 1995).

However the PSY and MAT groups exhibited some distinctive features. Within the PSY group, each stochastic item was categorized as not random by about a third of the subjects, while within the two other groups all stochastic items were categorized as random by almost all the subjects. For some psychologists an item is not random whenever it is possible to compute a probability. We assume that this marginal conception could be linked to their statistical practice. Indeed psychologists routinely use null hypothesis significance tests and a common presentation of this procedure is that rejecting the null hypothesis implies rejecting randomness and consequently could justify deterministic conclusions about the data. So, Tryon (2001) wrote "rejection of the null hypothesis implies that the results are not due to chance and that therefore they must be both systematic and reproductible." Furthermore psychologists categorized real items as not random more often than the other subjects.

A characteristic of the MAT group is that some subjects explicitly referred to two types of randomness: a "mathematical" randomness when it is easy to compute an objective probability (typically the stochastic items), and a randomness "when unknown" when it is not easy to compute a probability due to a lack of available standard probabilistic model (typically the real items).

Finally, it must be emphasized that the arguments used to judge if an event is random, and those to judge if it is not random, were found to be of different natures. In general subjects considered randomness to be involved in situations when probability was also involved, and considered randomness not to be involved when causal factors could be identified. To assess randomness, a large majority of subjects argued that "it is random because it is possible to compute a probability." All these subjects applied this probability based argument to the stochastic items, and approximately half of them to the real items that are consequently judged as random. It is interesting to note that, while the concept of probability has been introduced to formalize randomness ("randomness implies probability"), a majority of individuals appear to consider probability as a primary concept ("probability implies randomness"). By contrast only a minority of subjects referred to more direct arguments such as "it is random because one can't control or predict anything" ("no causality implies randomness"). To assess non randomness, the main argument is that "it is not random because there determinism plays a great part or because causal factors can be identified" ("causality implies non randomness"). Only a weak minority (two psychologists) used a probability based argument and surprisingly

argued that "it is not random because it is possible to compute a probability" ("probability implies non randomness").

It must be acknowledged with Shaughnessy (1992, p. 468) that "the model of probability that we employ in a particular situation should be determined by the task we are asking our students to investigate, and by the types of problems we wish to solve." Conversely, the types of problems should be enlarged to go beyond "stylized situations" in which probability assignments are essentially based on considerations of symmetry. They should, in particular, include situations involving probability judgments and predictions about single case events. Of course, this is not an effortless avenue. Clearly, we must not restrict our attention to the conventional frequentist view of probability. The Bayesian conception allows the students to assign a probability to a wider range of situations (Steinberg & Von Harten, 1982). Moreover, the Bayesian definition can be applied to real life uncertainty situations in which it is often not possible to "easily" compute a probability.

Concentrating more specifically on the teaching of the Bayesian statistical inference approach, instructors face the difficulty of explaining to students that the parameters, as well as the statistics (*before observations*), are considered as random. According to our results, this difficulty should be all the more serious in that it is not easy to assign a probability. On the one hand, the sampling probability distributions of statistics clearly refer to stochastic situations. At least, in most familiar situations, sampling probabilities are relatively easy to compute, and the level of mathematical justifications can be adapted to the students. On the other hand, Bayesian probabilities about parameters refer to single case events. The elicitation of the prior probabilities is precisely one of the most often denounced difficulties with the Bayesian approach. A possible approach, advocated by Berry (1997), is to place emphasis on the fact that prior and posterior Bayesian distributions are subjective and to force students to assess their prior probabilities. However , this task is not easy for the students and Berry recognized that "they don't like it." An alternative strategy, based on our teaching experience with Bayesian methods (Lecoutre, Lecoutre & Grouin, 2001; Lecoutre, 2006), is to avoid – at least *in a first stage* – the issue of assessing a subjective prior distribution and to focus the teaching on "objective Bayesian analysis" (Berger, 2004), based on *noninformative* priors. Such priors fit well with the conception of *randomness when unknown*. As for the sampling probabilities, the resulting posterior probabilities are relatively easy to compute and the level of mathematical justifications can be adapted to the students. Once students have become familiarized with their use and interpretation, the introduction of "informative" prior distributions *at a later stage* is generally well-accepted.

A considerable difficulty in the teaching of the frequentist approach is that data continue to be treated as random *even after observation*. This seems so *strange* to students that the frequentist interpretation of confidence intervals hardly makes sense for them. However, according to our results, it is not so paradoxical that most statistical users erroneously interpret the frequentist confidence level as the (Bayesian) probability of the single event "the parameter lies between two fixed limits." Indeed, since a probability is available for this event, these users have no doubt that it is a random event. Furthermore, all attempts to teach the orthodox frequentist interpretation seems to be "a losing battle" (Freeman, 1993). Our suggestion is to replace, as much as possible, probabilistic formulations about sampling distributions with formulations in terms of "proportions of samples." Thus, the probabilistic formulations are mainly reserved for the Bayesian approach, minimizing a possible source of confusion. In conclusion, it remains a challenge for statistics educators to reduce students' confusions about the different notions of probability. In this perspective, it is important that they become familiar with

the variety of meanings and beliefs about randomness. In particular, knowing what students think is, or is not, "random," in relation to probability based arguments, should facilitate the communication between students and teachers regarding probability and statistical inference. Our finding that background knowledge of probability theory has little effect on one's view of randomness implies that a mutual understanding is possible.

ACKNOWLEDGMENTS

Our special thanks go to Charlotte Détrie for improving our English. The remaining mistakes are ours.

REFERENCES

Albert, J. (2002). Teaching introductory statistics from a Bayesian perspective. In B. Phillips (Ed.), *Sixth International Conference on Teaching Statistics Proceedings.* Cape Town, South Africa [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute [Online: www.stat.auckland.ac.nz/~iase/publications/1/3f1_albe.pdf]

Albert, J. (2003). College students' conceptions of probability. *The American Statistician 57*, 37-45.

D'Agostini, G. (2000). Role and meaning of subjective probability: some comments on common misconceptions. *AIP Conference Proceedings*, *568*. Melville.

Barthélemy, J.-P., & Guénoche, A. (1991). *Trees and proximity representations*. New York: Wiley.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58-80.

Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 1-17.

Berry, D. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician*, *51*, 241-246.

Buneman, P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory (B)*, *17*, 48-50.

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition, 43*, 197-223.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

de Finetti, B. (1974). *Theory of probability* (Vol.1). New York: Wiley & Sons.

Freeman, P. R. (1993). The role of *p*-values in analysing trial results. *Statistics in Medicine*, *12*, 1443-1452.

Griffin, D., & Buelher, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology*, *38*, 48-78.

Hawkins, A. S., & Kapadia, R. (1984). Children's conceptions of probability – A psychological and pedagogical review. *Educational Studies in Mathematics, 15,* 349-377.

Jaynes, E. T. (2003). *Probability theory: The logic of science* (Edited by G.L. Bretthorst). Cambridge: Cambridge University Press.

Kac, M. (1983). Marginalia: What is random? *American Scientist*, *71*, 405-406.

Kadane, J. B. (1996). *Bayesian methods and ethics in a clinical trial design*. New York: John Wiley & Sons.

Konold, C. (1991), Understanding students' beliefs about probability. In E. V. Glasersfeld (Ed.), *Radical constructivism in mathematics education* (pp. 139-156). Amsterdam: Kluwer.

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

Konold, C., Lohmeier, J., Pollatsek, A. Well, A. D., Falk, R., & Lipson, A. (1991). Novices' views on randomness. In R. G. Underhill (Ed.), *Proceedings of the Thirteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 167-173). Blacksburg: Virginia Polytechnic Institute.

Laplace, P.-S. (1951). *A philosophical essay on probability* [English translation, original work published 1814: *Essai philosophique sur les probabilités*]. New York: Dover Publications.

Lecoutre, B. (2006). Training students and researchers in Bayesian methods for experimental data analysis. *Journal of Data Science*, *4*, 207-232.

Lecoutre, B., & Derzko, G. (2001). Asserting the smallness of effects in ANOVA. *Methods of Psychological Research*, *6*, 1-32. [Online: www.mpr-online.de]

Lecoutre, B., Lecoutre, M.-P., & Grouin, J.-M. (2001). A challenge for statistical instructors: Teaching Bayesian inference without discarding the "official" significance tests. *Bayesian Methods with Applications to Science, Policy and Official Statistics*, 301-310. Luxembourg: Office for Official Publications of the European Communities.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review*, *69*, 399-417.

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

Loredo, T. J. (1990). From Laplace to Supernova SN 1987A: Bayesian inference in astrophysics. In P. F. Fougere (Ed.), *Maximum Entropy and Bayesian Methods* (pp. 81-142). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Moore, D. S. (1997). Bayes for Beginners? Some Pedagogical Questions. In S. Panchapakesan & N. Balakrishnan (Eds.), *Advances in statistical decision theory* (pp. 3-17). Boston: Birkhäuser.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*, 330-357.

Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant.* [The origin of the idea of chance in children]. Paris: Presses Universitaires de France.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximities data. *Psychometrika*, *47*, 3-24.

Robinson, D., & Foulds, L (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131-147.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*, 319-345.

Savage, L. (1954). *The foundations of statistical inference*. New York: John Wiley & Sons.

Shaughnessy, J. M. (1992). Research in probability and statistics: reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics and learning* (pp. 465-494). New York: Macmillan.

Steinberg, H., & Von Harten, G. (1982). Learning from experience – Bayes' theorem: A model for stochastic learning. *Proceedings of the First International Conference of*

Teaching Statistics, Volume 2, (pp. 701-714). Sheffield, U.K.: Teaching Statistics Trust.

Thom, R. (1986). Preface of P.S. Laplace, *Essai philosophique sur les probabilités* [Text of the 5th edition, 1825]. Bourgois : Paris.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrate alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371-386.

Vranas, P. (2001). Single-case probabilities and content-neutral norms: A reply to Gigerenzer. *Cognition*, *81*, 105-111.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of the literature. *Psychological Bulletin*, *77*, 65-72.

MARIE-PAULE LECOUTRE

ERIS, Laboratoire Psy.Co, E.A. 1780

Université de Rouen, UFR Psychologie, Sociologie, Sciences de l'Education

76821 Mont-Saint-Aignan Cedex, France

http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris

ELEMENTARY PRE-SERVICE TEACHERS' CONCEPTIONS OF VARIATION IN A PROBABILITY CONTEXT

DANIEL CANADA
Eastern Washington University
dcanada@mail.ewu.edu

ABSTRACT

While other research has begun to contribute to our understanding of how pre-college students reason about variation, little has been published regarding pre-service teachers' statistical conceptions. This paper summarizes a framework useful in examining elementary pre-service teachers' conceptions of variation, and investigates the question of how a class of pre-service teachers' responses concerning variation in a probability context compare from before to after class interventions. The interventions comprised hands-on activities, computer simulations, and discussions that provided multiple opportunities to attend to variation. Results showed that there was overall class improvement regarding what subjects expected and why, in that more responses after the interventions included appropriate balancing of proportional thinking along with an appreciation of variation in expressing what was likely or probable.

Keywords: *Statistics Education Research; Teacher Education; Variation; Probability*

1. INTRODUCTION

The purpose of this paper is to report on research aimed at elementary pre-service teachers' conceptions of variation. Other research has already begun to illuminate pre-college student thinking about variation in several contexts, such as sampling (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999; Torok & Watson, 2000; Reading & Shaughnessy, 2004), data and graphs (e.g., Watson & Moritz, 1999; Meletiou & Lee, 2002; Reading, 2004), and probability situations (e.g., Truran, 1994; Shaughnessy, 1997; Shaughnessy & Ciancetta, 2002). Moreover, in keeping with the centrality of variation to the entire discipline of statistics (Cobb & Moore, 1997; Wild & Pfannkuch, 1999), an entire issue of the *Statistics Education Research Journal* focused on "research on reasoning about variation and variability" (Jolliffe & Gal, 2004).

As the picture begins to get painted about how pre-college students reason statistically, a continuing focus for research also needs to be on the teachers of these students: What sort of subject matter knowledge and pedagogical content knowledge do teachers have to prepare them for teaching and assessing in ways that enhance their own students' learning (Shulman, 1986)? Some researchers have incorporated inservice teachers into their studies, such as when Hammerman and Rubin (2004) looked at using statistical software tools in a professional development seminar and in the teachers' own classes. Garfield and Ben-Zvi (2005) begin to address the need to tie together the ways in which research informs practice as they present an epistemological model for teaching and assessing variability, but the research on how teachers reason about variation, or variability in data, remains thin.

Meanwhile, there is a paucity of research about how pre-service teachers think about variation. Makar and Confrey (2005) report on informal language used by secondary pre-service teachers to describe variation while reasoning about distributions, but little has been published regarding elementary pre-service teachers' conceptions of variation. Since university teacher preparation programs are concerned with both the subject matter knowledge as well as the pedagogical content knowledge of teacher candidates, it makes sense to attempt to identify the pre-service teachers' conceptions of variation. If a goal is for teachers to "provide students with authentic, inquiry-based tasks meant to develop children's reasoning about variation" (Makar & Canada, 2005), then a natural step in achieving this goal is to improve teacher training courses. By discerning components of pre-service teachers' reasoning, teacher educators can better design university experiences that promote an understanding of variation for pre-service teachers, as well as an understanding on how pre-college students come to learn this topic.

Therefore, doctoral research (Canada, 2004) was undertaken to explore which components of a conceptual framework help characterize elementary pre-service teachers' thinking about variation, how their conceptions of variation before an instructional intervention compared to those conceptions after the intervention, and what tasks were useful for examining their conceptions of variation in the contexts of sampling, data and graphs, and probability. Because so little is known about pre-service teacher knowledge in statistics education, the ultimate purpose of the research was to take first steps in discovering what pre-service teachers think regarding situations where variation was a key component. Since almost all of the subjects had no recollection of having ever taken prior courses that included any probability or statistics, their initial responses at the outset of the research may be considered intuitive. In that sense, the rationale behind the research was that by taking subjects who initially had little familiarity with the cognitive tasks being posed, and then studying how the thinking of those subjects changed from the more intuitive to the more substantive over the course of instructional interventions, key aspects of the subjects' thinking might be revealed.

In addition to summarizing the conceptual framework emerging from the doctoral study (within which the results of this paper are situated), this paper presents results for the following research question: How do elementary pre-service teachers' responses concerning variation in a probability context compare prior to and after class interventions? To address this question, first the overall methodology of the study will be presented, with particular emphasis on the structure of the class interventions and the nature of the survey and interview tasks. Then, the conceptual framework used to examine elementary pre-service teachers' reasoning about variation will be discussed, and connections will be made to past and recent models posited by other researchers. After presenting the research results and accompanying analysis, a summary and implications for teacher training programs follow.

2. METHODOLOGY

2.1. PARTICIPANTS AND DESIGN

The thirty subjects in the study of elementary pre-service teachers (24 women, 6 men) were enrolled in a ten-week pre-service course at a university in the northwestern United States. The course, Math for Elementary Teachers II (MET II), is designed to give prospective teachers a hands-on, activity-based mathematics foundation in geometry and probability and statistics. The only prerequisite course (MET I) focused mainly on whole and rational numbers and their operations. Thus, it was not expected that the subjects had

received any prior formal instruction in probability or statistics. When asked at the outset of MET II what classes involving probability or statistics they recalled taking (earlier in college or in high school), almost all subjects thought they had either not had any such classes, or they had taken such classes so long ago that they couldn't remember any specifics. For obtaining licensure to teach grades K-8, the two-course sequence of MET I and II represents the only content-specific math classes required at the university.

For the structure of the MET II course, Weeks 1 – 4 and 9 – 10 were on geometry and the four-week span from Week 5 – 8 was devoted to probability and statistics. During the first week of the course, subjects took an in-class survey (called a PreSurvey) designed to elicit their understanding on a range of questions about sampling, data and graphs, and probability. The PreSurvey probability questions reported on in this paper considered a hypothetical person Mark tossing a fair coin 50 times (a *set* of 50 tosses), with a focus on how many of the 50 tosses landed heads-up. The questions are given in Table 1.

Table 1. PreSurvey Q7 (Sets of 50 Flips of a Fair Coin)

Subquestion	Nickname	Description
Q7ai	One Set - What	How many times out of 50 flips do you think the coin might land heads-up?
Q7aii	One Set - Why	Why do you think this?
Q7b	Compare Sets	After Mark's first set of 50 flips, he decides to do a second set of 50 flips. How do you think his results on the second set of 50 flips will compare with the results of his first set?
Q7ci	Six Sets - What	Mark actually has a lot of time on his hands, so the next day he does 6 sets of 50 flips. Write a list that would describe what you think might happen for the number of flips out of 50 the coin would land heads-up (in each of the 6 sets of 50 flips).
Q7cii	Six Sets - Why	Why did you choose those numbers?

The PreSurvey also contained an invitation for subjects to participate in individual interviews after regular class hours. The purpose of the individual interviews was to have an open-ended time where subjects could expand on their views and provide deeper explanations than were possible on the surveys. Eleven subjects volunteered to be interviewed, and all eleven were scheduled for one-hour interviews (called PreInterviews) before commencing instruction in probability and statistics so that their conceptions of variation could be further explored. The interviews were videotaped and included some of the same questions that were on the surveys so that subjects' verbal responses could be compared with what they had written earlier and extensions to their thinking could be probed. Thus, the interview script contained specific questions that were used with each subject, but the protocol also allowed flexibility to follow each individual subject's unique train of thought.

During Weeks 5 – 8, a series of activities was conducted in class specifically designed to offer opportunities to investigate and discuss variation. The activities (comprising the Class Interventions) were centered on the three realms of data and graphs, sampling, and probability situations. The instructor of the course, Sam, allowed me to co-lead many of the activities in class with him. Take-home surveys (called PostSurveys) were given after each class intervention. The PostSurvey probability questions reported on in this paper considered a hypothetical person Matt spinning a half-black and half-white spinner 50 times (a *set* of 50 spins), with a focus on how many of

the 50 spins landed on black. The questions (Table 2) were isomorphic to those asked in the PreSurvey:

Table 2. Probability PostSurvey Q1 (Sets of 50 Spins of a ½ - Black, ½ - White Spinner)

Subquestion	Nickname	Description
Q1ai	One Set - What	How many times out of 50 spins do you think the arrow might land on black?
Q1aii	One Set - Why	Why do you think this?
Q1b	Compare Sets	After Matt's first set of 50 spins, he decides to do a second set of 50 spins. How do you think his results on the second set of 50 spins will compare with the results of his first set?
Q1ci	Six Sets - What	Matt actually has a lot of time on his hands, so the next day he does 6 sets of 50 spins. Write a list that would describe what you think might happen for the number of spins out of 50 the spinner would land on black (in each of the 6 sets of 50 spins).
Q1cii	Six Sets - Why	Why did you choose those numbers?

After shifting topics in the course from probability and statistics back into geometry for the final two weeks of the course, the same eleven subjects who participated in PreInterviews also participated in PostInterviews. In the PostInterview, subjects were asked to elaborate on their PostSurvey responses concerning the spinner tasks, using a protocol similar to that in the PreInterview.

While interview data were gathered from eleven subjects, only six representative cases were chosen from among those eleven for this paper. The reason for this selection was that the grounded-theory approach (used in discerning the aspects of the conceptual framework that was a main contribution of the research) enabled a point of saturation to be reached, beyond which new data was not adding anything new to the framework. Thus, taken cumulatively, the responses from the six interview subjects profiled in this paper may be seen as representative of the class as a whole.

2.2. CLASS INTERVENTIONS

In this section, the class interventions, comprised of activities around which much of the class discussion was based, are described in more detail These interventions are presented in the order in which they occurred in the MET II course, leading off with the context of data and graphs, then sampling, and finally probability situations.

Class Intervention #1 (Data & Graphs) The two activities comprising the Class Intervention for the context of data and graphs were called "Four Questions" and "Body Measurements." The first activity offered a good opportunity to discuss both average and spread in data sets, and Sam started the class exploration of statistics in the fifth week by having the entire class gather data from one another in response to four questions:

Four Questions Activity Prompt
How many pets do you have?
How many years have you lived in this city (to nearest half-year)?
How many people are in your household?
How much change (in coins) do you have today?

After graphing the data in different ways, the class had a discussion about levels of detail provided by each type of graph and about "typical" values for an individual student and for the whole class. The tension between centers and spread of data was one theme to emerge from the discussion of the graphs. For the second activity, everyone's own armspan, height, handspan, head circumference, and pulse rate per minute were recorded. Also, all students in class measured a designated person's armspan, to see how multiple measurements of the same object would compare. Again, we had a class discussion about the data and graphs for the body measurements, this time focusing more on causes of variation.

Class Intervention #2 (Sampling) In the seventh week of class, the two activities "Known Mixture" and "Unknown Mixture" were conducted with Sam's students. Prior to the "Known Mixture," we started with a general discussion of what samples were, who uses samples, and the purpose of sampling. Then the following scenario for the Known Mixture Activity was given as a part of a handout:

Known Mixture Activity Prompt

The band at Johnson Middle School has 100 members, 70 females and 30 males.
To plan this year's field trip, the band wants to put together a committee of 10 band members.
To be fair, they decide to choose the committee members by putting the names of all of the band members in a hat and then they randomly draw out 10 names.

The class discussed initial expectations for this scenario, focusing especially on what would happen if the random draw of 10 names were to be repeated thirty times. After students talked about predictions for drawing thirty samples each of size ten, we simulated this activity using chips in a jar. Actual data were gathered and graphed. Then we had a discussion about how the graphs of the predicted data compared to one another, how the graphs of the actual data compared to one another, and also how the predicted graphs compared to the actual graphs. We then made a transition into the second activity in this intervention, the Unknown Mixture. Now we had larger jars that each contained 550 yellow and 450 green chips, and the use of opaque jars having only a narrow opening made it difficult to look inside at the contents. Students were only given the information that each jar had 1000 total chips, and that the mixture was identical across all jars (they were not told the true mixture). To make a conjecture about the true mixture of chips, the students were asked to decide in their groups what sample size they wanted to use and how many samples they wanted to draw. Then they were to carry out their plans, do the sampling, graph the results, and make their conjectures about the true mixture in the jar. After the sampling was carried out, we had a class discussion about the different choices made in sampling and the class results, and we tried to forge a class consensus about what the true mixture was.

Class Intervention #3 (Probability) There were two activities that made up this intervention, "Cereal Boxes" and "The River Crossing Game." These were chosen specifically because of the probability aspects involved in the activities and these were the main class activites involving random devices. Cereal Boxes relies on the use of spinners and River Crossing on the use of ordinary fair dice as random generators.

Cereal Boxes actually took place in the first class session of week 6, just before we gathered data for Body Measurements. As explained earlier, there was considerable overlap in the three contexts, and Cereal Boxes is a good example of this overlap. Cereal

Boxes is a sample-until scenario, assuming that any of five different stickers can be obtained within each box of cereal, and that the five stickers have equal chances of being obtained. The question concerns how many boxes need to be opened to obtain all five stickers, and the situation was simulated by using an equal-area five-region spinner. Cereal Boxes brings together probability, sampling, and data and graphs in a way that highlights variation.

The second activity for this intervention, the River Crossing Game, involved finding the sum of the scores on two dice. Both the Cereal Boxes activity and River Crossing Game are part of the *Math and the Mind's Eye* curriculum (Shaughnessy & Arcidiacono, 1993). Using two players, each player receives 12 chips to place on their side of a "river," along spaces marked 1 through 12. After configuring their chips in an initial arrangement along the spaces, players take turns tossing a pair of dice. If either player has any chips on the space showing the total sum for the dice, one chip can "cross the river" and be removed from the board. The winning player was the first one to remove all the chips on his or her side. Note that although a sum of 1 is not possible to obtain, this fact is left for the players to discover; the challenge remains for players to reason about the optimal placement of their 12 chips. As with Cereal Boxes, in the River Crossing Game we made predictions, gathered and graphed data, and discussed results.

The activities in all the interventions were designed to elicit discussion about variation. For instance, the intervention on data and graphs included different types of graphs and the amounts of variation they showed. Body Measurements got at the ideas behind multiple measurements of the same object, whereas the Known and Unknown Mixtures had students actually draw chips from a container to experience variation resulting in a sampling context. Cereal Boxes and the River Crossing Game had students use traditional random generators such as spinners and dice to get a sense of what was likely in a probability context. Software, including ProbSim® (Konold & Miller, 1994) and Fathom™ (Finzer, 2001), was used to aid construction of graphical representations and to extend the simulations that the class had already participated in manually.

3. CONCEPTUAL FRAMEWORK

The doctoral study (Canada, 2004) had as one of its goals the description of components of a conceptual framework to help characterize elementary pre-service teachers' thinking about variation. Although different frameworks have been described by other researchers (e.g., Jones, Mooney, Langrall, & Thornton, 2002; Watson, Kelly, Callingham, & Shaughnessy, 2002), pre-service teachers were not the subjects of such research. Thus, the approach used in this study was largely exploratory, relying heavily on grounded theory, to let the data provided by the subjects help fill in details of what was called an *Evolving Framework.* Data which helped develop the framework included the written PreSurvey and three PostSurveys, transcriptions from all of the PreInterviews and PostInterviews, and observations from the class interventions.

The framework provides a lens through which three different *aspects* of an elementary pre-service teacher's understanding of variation can be viewed. The three aspects address how subjects reason in terms of *expecting*¸ *displaying*, and *interpreting* variation; these aspects are then defined in terms of their constituent *dimensions* (Figure 1). The following subsections describe the dimensions of the framework in terms of major themes that emerged from the participants in the study, and connections to framework components posited by other researchers are also discussed.

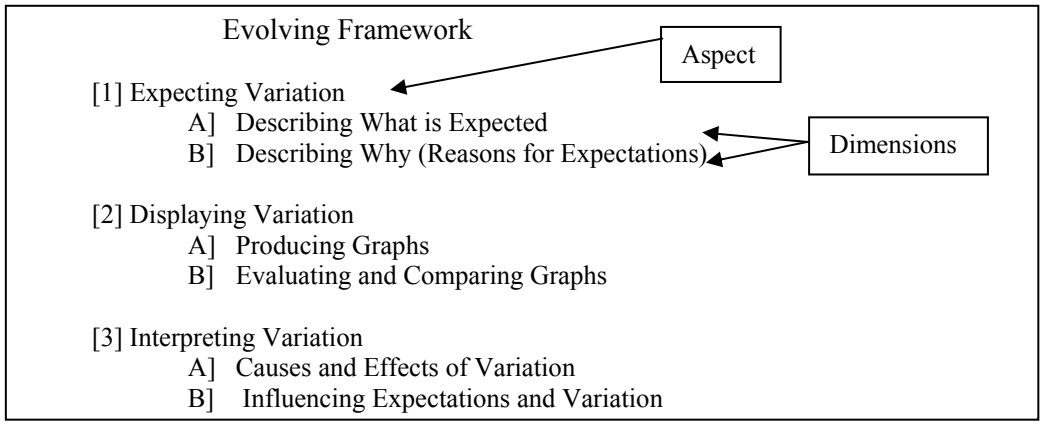

Figure 1. Framework for elementary pre-service teachers' conceptions of variation

3.1. EXPECTING VARIATION

When expecting variation, such as prior to sampling or conducting probability experiments, subjects expressed both *what* they expected and *why*. The expected value or average was a frequent theme concerning *what* subjects thought might occur in predicting results for experiments involving uncertain outcomes. A dominant type of response was how results should be close to, about, or near the expected value, and a more explicit type of response was how results might be higher or lower than the expected value. Another theme for *what* was expected concerned ranges or extreme values. More specific than just suggesting that results be above and below the expected value, some responses actually specified a numeric range.

In describing *why* they held their expectations, almost all subjects' reasoning at some point involved the language of possibilities and likelihood. For example, many subjects explained how extreme results were possible but unlikely. Reasons involving personal or shared experience constituted another theme. Some subjects mentioned informal out-of-class experiences, such as games they had played, and other subjects recalled their experiences with the in class activities. The theme of proportional reasoning can be a useful anchor to help center expectations appropriately, and this theme was a part of many subjects' reasons for *why* they expected what they did. An over-reliance on proportional reasoning can lead to a restricted expectation of variation, but an under-reliance on proportional reasoning can also lead to poor expectations. Some subjects were less influenced by proportional reasoning than by additive reasoning, as in the case of sampling experiments where the quantities in a sample or a population were more persuasive than the proportions.

3.2. DISPLAYING VARIATION

Subjects showed their skills and reasoning along the two dimensions of *producing graphs* as well as *evaluating and comparing graphs*. In considering how subjects *produced* graphs for tasks such as predicting the results of 50 samples of size 10 taken from a jar of 60 red and 40 yellow candies, the technical details of their graphs were a reflection of the subjects' own graph sense. Subjects often drew smooth bell curves in situations when a bar chart or dotplot would have been a better choice. Some graphs had detailed axes with an appropriate scale, while others had unlabeled axes or inappropriate scales. How the characteristics of the distribution get conveyed is another theme. Subjects

generally gave centers that were reasonably placed, but they often provided ranges that were too wide. Spreads were occasionally too tight or too scattered, and shapes were often unnaturally symmetric.

When *evaluating and comparing graphs,* such as comparing average annual traffic rate deaths between two regions of America, the four themes exhibited by subjects' responses corresponded to four components of distributional reasoning: Average, range, shape, and spread. A focus on average, was reflected in most but not all the subjects' responses. Many subjects were able to move beyond a focus on average to include references to other features of the distribution, but some made it clear that the average was their primary consideration in answering any question having to do with graphs. The theme focusing on range or extremes was often reflected in questions having to do with which graph had more variation. Subjects had some standard ways of talking about shape, using language like "symmetric" or "skewed," "normal" or "uniform." There were also some non-standard ways of referring to the shape of a distribution, including the use of hand gestures to try to communicate the picture in the subject's mind. Responses that focused on the theme of spread depended on the type of display that subjects were considering. For example, when using dotplots, some subjects referred to the way data were "clustered" or "scattered" along the horizontal axis to indicate how they saw the way data were grouped or spread out in the graph. In using boxplots, class discussions had focused on how the interquartile range was one measure of spread, and many subjects referred to that measure in their responses.

3.3. INTERPRETING VARIATION

Two dimensions that arose in the data for this aspect were *causes and effects of variation*, and *influencing expectations and variation*. Under *causes and effects of variation*, one theme reflected in the data was naturally occurring causes, such as the geographical and meteorological causes subjects listed for differences in rainfall patterns between two cities. In the sampling and probability situations, many students seemed to point to randomness as a naturally occurring cause. The other theme of physically induced causes included those causes which were deliberate or intentional as opposed to naturally occurring. For example, lining up the spinner in the same spot for each spin and trying to apply the same amount of force each time was seen a physically induced cause for reduced variation. The *effects of variation* were seen in terms of two distinct but related themes: the effect of variation on students' perceptions and the effect of variation on their decisions. For example, some students perceived a difference between theoretical predictions and real-life outcomes, and many students perceived that "anything can happen" in situations involving variation. Also, some students expressed a lack of confidence in making decisions, reflected in their "I don't know" responses. In making inferences, it seems that the two themes for *effects of variation* were often linked. For example, a student who thinks that "anything can happen" may be thinking that there is no way to decide what might happen, and thus the student may respond with "I don't know."

The two themes for *influencing expectations and variation* were quantities in sampling (i.e., the numbers of candies in the population or in the sample) and also the numbers of samples taken. The first theme applied primarily to the context of drawing samples where there was a discrete population, such as samples of candies from a jar containing 60 red and 40 yellow candies. Several subjects focused on the sheer numbers of candies in the jar, and in some cases it seemed that the probabilities of getting different outcomes were linked to these quantities. Particularly for subjects who are not strong

proportional reasoners, there may be a tendency to see the quantity and not the ratio as the influential factor in the behavior of the sample outcomes. The second theme, involving the numbers of samples taken, was reflected in many different ways. Almost all of my subjects pointed out that more samples would widen the overall range, while very few subjects suggested that more samples would also tighten the subrange capturing most of the results. Other ideas included how additional samples offered more chances to attain the expected value, and how additional samples provided a better picture of the underlying distribution.

3.4. RELATION TO OTHER MODELS

The summary of the Evolving Framework provided in this section captures some of the main ways in which subjects expressed their intuitive and emerging conceptions of variation throughout the doctoral study (Canada, 2004). Grounded in survey, interview, and classroom observation data, the framework provides structure for characterizing elementary pre-service teacher thinking about variation in the contexts of sampling, data and graphs, and probability situations. The framework is "evolving" because there are no doubt more ways in which elementary pre-service teachers' understandings of variation can be modeled, and the framework is expected to grow as more comparisons to other models of thinking are made. Already the aspects of the evolving framework reflect facets of other models. For example, Wild and Pfannkuch (1999) incorporated acknowledging, measuring, modeling, and explaining variation within their components of a model for statistical thinking. Acknowledging variation is involved when explaining what is expected regarding variation, and also relates to producing, evaluating, and comparing graphs when dealing with displays of variation. Explaining variation relates both to explaining why people expect what they do and also to causes of variation. To the model of Wild and Pfannkuch, Reading and Shaughnessy (2004) added the two components of describing and representing variation. Reading and Shaughnessy's description hierarchy reflected what was expected in terms of extreme and central values, and also how expectations deviated from an anchor. Also, a causation hierarchy included extraneous (physical) causes of variation as well as the reason why results might vary, such as additive or proportional reasoning. More recently, Garfield and Ben-Zvi (2005) proposed the following seven dimensions of a theoretical framework representing key facets of understanding variation, or variability in data:

(1) Developing intuitive ideas of variability
(2) Describing and representing variability
(3) Using variability to make comparisons
(4) Recognizing variability in special types of distributions
(5) Identifying patterns of variability in fitting models
(6) Using variability to predict random samples or outcomes
(7) Considering variability as part of statistical thinking

The framework proposed by Garfield and Ben-Zvi (2005) provides a comprehensive structure for looking at how people reason about variation and incorporates multiple aspects of other researchers' models of conceptualizing variation. For the evolving framework looking at elementary pre-service teachers' conceptions, certainly their intuitive ideas were explored in terms of what variation they expected and why. How elementary pre-service teachers dealt with displays of variation addressed the ways in which they described and represented variation, and also how they used variation to

compare distributions. Subjects also had primitive ways of using variation in making predictions for what they expected when sampling or considering probability outcomes.

While the evolving framework had as its three aspects *expecting, displaying,* and *interpreting* variation, the focus of this paper is on survey and interview results for a small subset of questions, namely the PreSurvey questions on coin flipping and the analogous PostSurvey and PostInterview questions concerning half-black and half-white spinners. These questions (presented earlier in Tables 1 and 2) did not encompass displays of variation, so the aspects of the framework that situate the results for this paper are how elementary pre-service teachers *expected* and *interpreted* variation.

4. RESULTS AND ANALYSIS

Results are first presented showing how (as a class) subjects' written survey responses concerning variation in a probability context compare before and after the class interventions. The comparison is facilitated by a scoring rubric derived from previous research and relating to the conceptual framework. Assessing paired-data results according to the rubric bolstered the claim of overall class improvement from the PreSurvey to the PostSurvey. Then, in the second and third subsections, results from both survey and interview questions are used to focus more sharply on changes in subjects' thinking according to two aspects of reasoning about variation in a probability context that emerged. These two aspects concern the subjects' expectations and interpretations of variation. Thus, the first subsection provides more of a quantitative backdrop for looking at overall class shifts in thinking, while the second and third subsections give more of a deeper, qualitative picture of elementary pre-service teachers' thinking about variation in accordance with the conceptual framework for the study.

4.1. CLASS SURVEY PERFORMANCE

To compare class performance on questions from the two surveys, coding schemes were adapted from rubrics used in a similar set of questions involving sampling candies out of a jar (Shaughnessy, Ciancetta, & Canada, 2004). After assessing the PreSurvey responses using the coding schemes, I had two colleagues (who were familiar with the original sampling rubrics) independently assess the responses. There was an initial inter-rater agreement of 91% on the PreSurvey, and all disagreements were subsequently resolved.

What follows are the class results for the PreSurvey and PostSurvey probability tasks described earlier, organized according to the task subquestions. Although class enrollment was 30, there were three absences on the day of the PreSurvey, which was completed during class time. After all the class interventions had taken place, the PostSurvey for Probability was the final written research instrument given in class, and was completed by 29 of the 30 enrolled students. The percentages are given of students who were coded at each of the levels for the different subquestions, and example responses from both surveys are also provided. Then changes in overall class performance on the different subquestions are discussed.

One Set The codes and class results for the first part of this subquestion are presented in Table 3:

Table 3. Results for One Set – what (PreSurvey Q7ai & PostSurvey Q1ai)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
2	Either gives a range around 25 such as 22-28, or else writes (for example) "Around 25."	1 (3.7%)	10 (34.5%)
1	Gives only 25 as answer.	21 (77.8%)	14 (48.3%)
0	Gives one number other than 25, such as 23.	5 (18.5%)	5 (17.2%)

As has been noted with similar sampling tasks involving predicted outcomes for one trial, most students put down the expected value (Shaughnessy et al., 2004), which in this case is 25. However, the number of students who volunteer some form of variability in their Level 2 response increased from PreSurvey to the PostSurvey. In fact, the average coding levels for class performance on this subquestion for both surveys go from 0.85 on the PreSurvey to 1.17 on the PostSurvey.

Even stronger evidence of class improvement is offered by the reasoning component to the subquestion, where subjects described why they held their particular expectation, and Table 4 has the codes and class results for the second part of this subquestion.

Table 4. Results for One Set - why (PreSurvey Q7aii & PostSurvey Q1aii)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
3	Uses proportional reasoning with some explicit statement about what else might happen.	2 (7.4%)	8 (27.6%)
2	Uses proportional reasoning (for example: ratio, average, or percent).	17 (63.0%)	17 (58.6%)
1	Uses additive reasoning, or gives a reasonable response which makes sense but lacks specificity.	4 (14.8%)	4 (13.8%)
0	No reason, a vague reason which makes no sense, or an irrelevant reason.	4 (14.8%)	0 (0.0%)

Here are some examples of responses from each coding level:

[Level 0]

Sarah (Q7aii) Maybe a little more than half 'cause it started on heads: I have no idea really.

[Level 1]

Rosie (Q7aii) Because you have the same chances.

Ross (Q1aii) There is no reason to expect that either white or black would result any more than the other, but an exact result isn't possible to predict.

[Level 2]

Emma (Q7aii) He has a 50% chance of landing on heads.

Brita (Q1aii) Because there is a 1 out of 2 (or 50%) chance that he will get black. So theoretically half of the spins will be black.

[Level 3]

James (Q7aii) The coin has a 1:2 chance of landing on heads. The more often you flip, the chances of the 1:2 ratio will be closer to that – 1 in 2.

George (Q1aii) There is a 50% chance of landing on white or black, in the long run it balances out closer and closer to 50%, but the short run it varies wider. 25 would be 50%, so 18 is very probable. It could be 28 or 20 or 16, but the more times, the closer it will be toward 25.

Additive reasoning was initially conceived as a level in connection with sampling tasks, where prior research had shown subjects to focus on the sheer numbers used in the sample or population (as opposed to a consideration of the proportion). In transferring the idea of additive reasoning to probability tasks, I had anticipated a focus on the amount of shaded area on a spinner, for example, or the number of sides for the coin. Some students did comment on the amount of shaded area in dealing with other spinners such as a 1:3 white-to-black spinner, but they did not use the same language for the 1:1 spinner used in this question.

Overall, there was a lack of specificity in the Level 1 responses, such as when Rosie wrote about the "same chances" without identifying what those chances were. Level 2 responses were characterized by the use of percentages, odds, or ratios, but little other information was usually given. In James' Level 3 PreSurvey response, he uses the ratio defined by the fair coin, but his response also suggests that the cumulative average of many flips approaches that ratio. Thus he shows thinking that aligns with the Law of Large Numbers, as does George's Level 3 PostSurvey response. Although George's written comment about "...the more times, the closer it will be toward 25..." doesn't make it clear what he is thinking about, in subsequent interviews it became apparent that he was envisioning the proportion of heads gravitating toward the theoretical 50% with increasing numbers of flips. One noticeable feature of Table 4 is how more students gave a Level 3 type of response in the PostSurvey than in the PreSurvey. Also, the average coding levels for class performance on this subquestion for both surveys go from 1.63 on the PreSurvey to 2.14 on the PostSurvey, again showing a sizable increase.

Compare Sets For this subquestion, a key idea to learn from responses was whether or not subjects believed that results on a second set of flips or spins would or should match the results from the first set. In sampling tasks used on the PreSurvey and PostSurvey and by other researchers (Reading & Shaughnessy, 2004; Shaughnessy et al., 1999; Shaughnessy et al, 2004), a similar question asked was "If several samples were taken, do you think you'd get the same results each time?" Subjects who were unduly influenced by the expected value often did answer affirmatively, with the idea being that if the expected value was reasonable for a single sample, then that same value was reasonable for several samples. In one study of 188 high school students, 25% agreed that results should be the same every time (Shaughnessy et al., 2004).

On the probability subquestions using flips and spins, the wording was changed (from what had been used on the sampling question described in the previous paragraph) so that subjects were invited to describe how results on a second set of 50 flips or spins might compare to the first set. The reason for the change in wording was to allow more flexibility in how subjects responded, since "...do you think you'd get the same results each time?" seemed to invite a straightforward yes-or-no kind of response. One thing to take into account with the less straightforward responses that the wording of the *Compare Sets* subquestions invited was that subjects often used "similar" to mean "similar but not

the same." Thus, the terms "different" and "similar" both occurred to signify "not the same." Classifying responses at different levels involved looking at other information provided showing *what* subjects expected and *why*. The codes and class results for this subquestion are presented in Table 5:

Table 5. Results for Compare Sets (PreSurvey Q7b & PostSurvey Q1b)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
3	[Different or Similar] with Explicit mention of a range or spread.	3 (11.1%)	15 (51.7%)
2	[Different or Similar] with Some additional information, such as use of ratio, average, percent, or giving specific alternatives for results.	11 (40.7%)	10 (34.5%)
1	[Different or Similar] with No additional information provided.	7 (25.9%)	4 (13.8%)
0	Mentions how results will be the same.	6 (22.2%)	0 (0.0%)

Over 20% of the PreSurvey responses indicated results would be the same, with responses such as:

[Level 0]

Ross (Q7b) In the absence of any change of approach, the results are most likely to be the same.

Maya (Q7b) Same. Probability will remain the same.

Ross's response points more to the physical aspects of doing the coin flips, implying that if the coin is flipped in the same manner, then obtaining the same results is "most likely." Maya shows she is influenced primarily by the constancy of the theoretical probability. No responses in the PostSurvey expressed a sense of expectancy that identical results would occur.

Level 1 responses only included an expression of difference or similarity, such as:

[Level 1]

Sally (Q7b) They will be similar but not the same.

Julie (Q7b) Similar, though probably a little different.

Jackie (Q1b) Could be slightly different, but basically the same.

Susie (Q1b) Probably very similar to the first set of results, keeping in mind that it is 'chance.'

Sally's response lends credence to the assumption that "similar" connotes "not the same," and judging by that assumption, most of the class in the PreSurvey (and all students during the PostSurvey) held the idea that results on the second set would likely not be identical to the first set.

Level 2 responses added additional information about what might happen or why:

[Level 2]

Carrie (Q7b) Probably different. But still has a 50/50 chance.

Emma (Q7b) He might get a few extra tails-up so his results should vary.

Cassie (Q1b) The comparison should be somewhat the same. It has the same odds again, 50%.
Robbie (Q1b) I think they will be very similar to the first set of 50 spins because the probability of getting black remains ½.

Whereas Emma clearly indicates an expectation of variation in results, what set apart the Level 3 responses was an explicit statement of *what* variation might result, or *how* results might vary:

[Level 3]
Maria (Q7b) It will be nearly the same, or the same. The variation may be only 2-3 one way or the other.
George (Q7b) Could be 30, 25, 20, 27.. .If he was super super super super lucky he'd get 50.
Molly (Q1b) Maybe a little different but still somewhere around 20-30.
Sofia (Q1b) Similar. Maybe a little wider range, 18-32.

Over half of the PostSurvey responses were at Level 3, suggesting that the class interventions helped attune students to thinking in terms of a range of expectations. In terms of class averages, again there was an increase in means, from 1.41 on the PreSurvey to 2.38 on the PostSurvey.

Six Sets Both parts of this subquestion (the *what* and the *why*) were taken into consideration for coding purposes, primarily to retain consistency with the analogous rubric derived for the similar questions in a sampling context (Shaughnessy et al., 2004). Only inappropriate choices for listing *what* was expected (or blank answers) were coded at Level 0. Deciding what would constitute an appropriate choice for the results on six sets of flips or spins involves making a judgment call, and the subcodes used for this subquestion question help identify inappropriate choices as (W)ide, (N)arrow, (H)igh or (L)ow. Of key interest was how many subjects had a narrow response consisting of just a list of six identical values, namely the expected value of 25. In research involving 93 high schoolers and a sampling task, almost 26% of responses were narrow, which was conjectured to be because of "an influence of probability instruction, or just lack or exposure to statistics tasks involving variability" (Shaughnessy et al., 2004, p. 6). The codes and class results for this subquestion are presented in Table 6.

Table 6. Results for Six Sets (PreSurvey Q7c & PostSurvey Q1c)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
3	Appropriate choice & Explanation explicitly involves proportional reasoning as well as variation.	2 (7.4%)	9 (31.0%)
2	Appropriate choice & Explanation reflects proportional reasoning or notions of spread.	10 (37.0%)	15 (51.7%)
1	Appropriate choice & Explanation left blank or lacks any specific reasons relating to details of the distribution.	4 (14.8%)	3 (10.3%)
0	Inappropriate choice (Regardless of Explanation). W(ide) = Range > 19, N(arrow) = Range < 2, H(igh) = Choices > 24, L(ow) = Choices < 26	11 (40.7%)	2 (6.9%)

Of the eleven inappropriate PreSurvey responses, one was narrow, one was high, one was low, and four were wide (the remaining were left blank). It was clear from subsequent discussions in class that students initially felt uncomfortable venturing a guess for six results, often demonstrating that it was difficult to guess correctly. Such an attitude toward expectation has much in common with the *Outcome Approach* to random events, whereby subjects look at the goal of probability as correctly determining ahead of time what will be the next outcome (Konold, 1989). Of the two inappropriate PostSurvey responses, both were wide.

A few of the Level 0 examples are:

[Level 0]

Alice (Q7c) {25, 25, 25, 25, 25, 25} I don't see how the chances of getting heads will change if he does more sets of 50 flips.

Brita (Q7c) {7, 21, 23, 25, 29, 31} I chose numbers close to 25 because I think with a 50% probability, the results would come out pretty close to 25. I put the oddball 7 in for fun, because there is always that element of chance.

Susie (Q1c) {5, 15, 30, 40, 45, 50} It is chance.

Alice's narrow response is obviously over-influenced by the expected value, but it seems surprising that more subjects did *not* put all 25s for their choices in the PreSurvey, given results discussed by other researchers (e.g., Shaughnessy et al., 1999). Brita's choice of 7 is extremely unlikely and makes her overall range too wide, although her upper bound of 31 is plausible). Susie's choices are too extreme at both the upper and lower ends. Level 1 responses had appropriate choices for *what* was expected but the reasons *why* did not specifically reflect distributional thinking:

[Level 1]

Carrie (Q7c) {22, 23, 24, 25, 26, 27} It's usually not the same.

Maria (Q1c) {20, 23, 25, 25, 26, 30} I think he will hit 25/50 one time. The rest of the times, he will be close, but not exactly on. Also I think he will be controlling the way he hits the spinner more on the second day, which accounts for no 23 or 28.

Maria points to *causes* of variation in noting the physical manipulation of the spinner, and other subjects also seemed to indicate that spinners are not viewed as true random devices because the user can ostensibly control outcomes by altering the way the pointer is spun. The Level 2 responses included an indication of reasoning using an average, proportion, or a measure of spread:

[Level 2]

Sofia (Q7c) {20, 20, 24, 25, 26, 27} Because they average to about 25.

Sally (Q7c) {22, 23, 24, 25, 26, 27} They are all close to 25, ½ of 50.

Leila (Q1c) {23, 24, 24, 25, 25, 26} The numbers are pretty close to half or 50%.

Rocky (Q1c) {20, 22, 23, 27, 28, 30} These numbers represent a distribution across range of likely results.

Note how Sally's Level 2 response includes the same choices as in Carrie's Level 1 response shown earlier. However, Sally gives more specificity than Carrie in describing

her reasoning, which is proportional in Sally's case. Rocky doesn't include the expected value in his choices, but feels he has given a likely range and his sophisticated language borders on a Level 3 response. What distinguished the Level 3 responses was an indication of reasoning using *both* centers and spread:

[Level 3]

Maya (Q7c) {23, 24, 25, 25, 25, 26} Because there should be variation around the mean. The average should be 25.

Ross (Q7c) {22, 23, 24, 26, 27, 28} While 25 flips are likely to be heads, in reality some variation is likely, so my numbers represent a range that averages 25.

Sally (Q1c) {21, 24, 25, 26, 28, 29} All numbers are 25 or close to 25 (1/2 of # of spins). Not all are 25 in order to account for variation.

Daisy (Q1c) {18, 21, 24, 26, 28, 31} Because they are close to the 50% chance to get 25 hits of black allowing for variation due to random spinning hits. But none of the #'s are too high or too low (far from the 25) which would be hard to hit based on the 50% odds.

There were more Level 3 responses in the PostSurvey than in the PreSurvey, and the relative sophistication is apparent as subjects reconcile the tension of having results close to an average value while also acknowledging the presence of variation. Also, there were more subjects in the PostSurvey than in the PreSurvey whose choices did not include the expected value of 25 (such as Susie, Rocky, and Daisy), suggesting that the class experiences helped counter the natural tendency to pin expectations solely to a theoretical average without an appreciation of the variation in subsequent trials. As in the other subquestions, average class performance on *Six Sets* also increased, from a mean of 1.11 in the PreSurvey to 2.07 in the PostSurvey.

Further Analysis Although the PreSurvey and PostSurveys were completed individually, the entire structure of the class interventions was geared towards small-group and whole-class discussions. With the constant exchange of ideas, opinions, and explanations that went on throughout the course, it made sense to look at classwide changes from the PreSurvey to the PostSurvey. On each subquestion, average class performance increased, with more students being rated in the highest coding level for each subquestion's scoring rubric.

Although there were 27 subjects who took the PreSurvey and 29 who took the PostSurvey, there were 26 subjects who took both. Thus, *t*-tests for differences in mean scores were applied to the paired data, with μ_1 as the mean for the PreSurvey and μ_2 the mean for the PostSurvey.

Using a one-sided test with H_a: $(\mu_2 - \mu_1) > 0$, and a significance level of $\alpha = 0.05$, statistically significant gains were found in all subquestions. Table 7 contains the p-values associated with each subquestion.

4.2. EXPECTATIONS OF VARIATION

Having seen evidence of classwide changes on the surveys, the following subsections focus on situating these changes more fully within the aspects of the conceptual framework. To help describe key aspects of understanding variation which emerged from the subjects, and how their expectations and interpretations changed throughout the course, examples of thinking provided by six interview subjects (Ross,

Table 7. Paired-data results for difference of means

Subquestions	Nickname	$\mu_2 - \mu_1$ Mean (Std. Dev)	*t*-test (25 df) p-values
Q7ai & Q1ai	One Set – What	0.27 (0.67)	0.0250
Q7aii & Q1aii	One Set – Why	0.46 (1.03)	0.0150
Q7b & Q1b	Compare Sets	0.96 (1.25)	0.0003
Q7c & Q1c	Six Sets – What & Why	1.00 (1.30)	0.0003

George, James, Daisy, Emma, Sandy) will be used along with examples from the class surveys and class discussions.

Describing What is Expected An important change from before to after the class interventions was how initial expectations that were overly influenced by the expected value became tempered by an increased appreciation for how variation occurs in multiple trials. Many students initially were inclined during the PreSurvey and PreInterview to think that the theoretical expected value was what *should* happen on any given trial (whether a sample drawn from a population or a set of flips or spins). These students also thought that even if results varied, the average of results *should* be the expected value. Such a perspective reflects the Law of Small Numbers described by Kahneman and Tversky (1972). By the end of the course, many students were speaking more about expectations in terms of a range rather than in terms of a single given value. For example, on the PreSurvey all six interview subjects expected 25 for the *One Set* of 50 flips, pointing out how 25 was the expected value. But for the *One Set* of 50 spins, five of the cases described their expectations on the PostInterview in terms of ranges. Ross expected "somewhere between 21 and 29...I would say – it's probably within that range," and James thought the result would be "approximately" 25, adding that "it will be, you know, plus or minus, maybe, 20% of that number – Somewhere in there." Daisy talked about the result being "within 2 or 3" of the expected value, and Emma and Sandy both said it would be "between 20 and 30 spins" for the result of black. Moreover, on *Compare Sets* in the PreSurvey, Ross felt the results on the second set "most likely were the same" as on the first set, whereas he had a different idea on PostSurvey, writing that the second set would likely show "not the exact same results." As he explained during the PostInterview, "I think that it's likely to fall in a same range, similar range." Sandy's comment was that "I think the range would still be somewhere very similar to that (first) one."

Another trend in responses was to avoid repeating choices when making predictions for multiple trials in the PostSurvey and PostInterview. For example, in giving choices for *Six Sets* on the PreSurvey, most students gave some repeated values for their choices, such as James' (20, 22, 25, 25, 26, 27) or Daisy's and Emma's (23, 24, 25, 25, 26, 27). There's nothing wrong per se with having repeated values in six conjectured results, but it is interesting how all of the interview subjects (and most of the other class members) had PostSurvey choices for *Six Sets* that contained no repeated values. Ross's choices on the PostSurvey *Six Sets* were (22, 23, 24, 26, 27, 28), and in the PostInterview he amended those to have a slightly wider range (21, 22, 23, 27, 28, 29), pointing out how his choices still were "similar, but not identical" and how "there's no repeats." Sandy said of her PostSurvey choices (20, 23, 24, 26, 28, 29) that "they could repeat, but I just did a

range – from 20 to 30, just to choose... different numbers, but still somewhere in that range." Ross and Sandy, like many others, also seemed to deliberately avoid including the expected value among their choices in the PostSurvey.

Describing Why (Reasons for Expectation) Improvements in reasoning for *why* students held their expectations included an emphasis on proportional reasoning combined with an understanding of what was probable or likely in the face of variation. Reading and Shaughnessy (2004) placed likelihood based proportional reasoning at the top of their causation hierarchy, which was related to why students gave their responses. In responses from the PostSurveys and PostInterviews, repeated results were unexpected because they were seen as unlikely, and extreme values were often described as unlikely but possible. Subjects also used probabilistic language in a general way, for instance talking of how the chances for events were seen as high or low, or about what might or could happen. For example, note the subjective use of language as Daisy reasons in the PostInterview about getting or not getting the expected value of 25 in *One Set* of 50 spins:

> Daisy: 50% would be 25, and I think it'd be rare that we'd get exactly 25 on our first spin. Well, not rare, but unusual. I mean, it's possible, but I think probably your first set of 50, it would be unusual that we'd get exactly 25 blacks. There's no guarantee that... you're going to get exactly 25 out of 50.

For multiple sets, she thought that "to get every single set of spins to be 25 would just be unlikely," and George suggested that results "could be higher, it could be lower...to not get a 25, it's possible that's not happening" Similar reasoning was used when discussing extreme results resulting from *Six Sets*:

> Emma: Sometimes you CAN get as low as a 16, and sometimes you can get as high as 34... It just seemed out of six (sets) that it's unlikely to get 45.
> George: Well, 25 would be half, and 20 is possible. It's possible to get a high number. You know, is it possible to get 36? Could it happen? Sure! Sure it could happen. It's very unlikely.
> James: I thought it was kind of unlikely that out of (six sets), to have a 10 and a 45...they just seemed too far out. Very unlikely in six (sets), but – Possible.
> Ross: (Commenting on a conjectured range) You've got a range here of 19 to 32, so it's hovering around that 25, and there is some variation but it doesn't strike me as extreme, and so… It seems possible, reasonable.

The examples provided by the interviewed subjects reflect the trend shown by most of the rest of the class to talk more in terms of what was likely or unlikely than in terms of what would or would not happen. Having had students become more sensitive to presence of variation, they were less strident in their predictions than they were at the beginning of the course, making them more cautious in their predictions. Many simply pointed to the presence of variation in their explanations, such as when Daisy, James, and Sandy commented about results for *Six Sets*, saying:

Daisy: 'Cause there's gonna be variation, because the spinner CAN land anywhere, but probably on average it'll be close to 25. You have variation from your 50%, a little variation from the 25, but not too much!
James: Well, he's just going to have some variation, even though … We know that the probability is 50%.
Sandy: Again, you wouldn't expect to get the same exact thing, I expect more variation.

Giving the presence of variation as a reason for an appropriate distribution of results, coupled with the use of proportional reasoning, and couching explanations in terms of possibilities and likelihoods were all indicators of improved reasoning about expectations.

One hypothesis about why so many students in the PostSurvey gave expectations in terms of ranges and choices without repeat values, and often even stayed away from including the expected value, is that the class experiences and discussions began to persuade them of the rather extreme unpredictability inherent in small numbers of repeated trials. For instance, in the River Crossing Game, some students talked about how they knew a sum of seven was the most likely outcome for the sum of two dice. Eventually the whole class knew the theoretical probability of a sum of seven for a pair of dice, namely $^1/_6$. However, it was clear that even if we threw the pair of dice six times, we might not see any sum of seven. Similarly, it became rather unremarkable in class to toss a fair coin ten times and *not* get exactly five heads, or spin the 5-Spinner in Cereal Boxes ten times and *not* get exactly two 1s as theory suggested. Through discussion and experimentation, students became more comfortable with ranges than with point estimates, and less comfortable with just sticking with the expected value in making predictions. Certainly many students discussed the influence of the class activities in explaining *why*, such as when Emma justified her prediction for *One Set* by recalling that "from doing the activity in class, I know it won't be exactly 50% but somewhere close." Other examples show the scope of the class comments about class activities:

Dixie: In our class experiments, I found when I repeated an experiment you'd often have some new variations pop into the picture but the central probability remains the same.
Rosie: Because we had the same activity in class, the same concept. I think that as we practiced in class, the more chances or tries you have the more different answers you can get.
Taha: Because due to the data shown in class, the majority of the data will be in the middle but there will be more variety with more data.
Sergio: I choose this answer judging my prediction on the exercise done in Monday's class because as demonstrated in class, every (trial) is different.

A particular impression was made by the computer simulations (using ProbSim and Fathom) that we did as a class, whereby we had a class discussion even as we continually ran the simulation with more and more trials. In later comments on the likelihood of getting extremes results, subjects clearly recalled the use of the computer:

Emma: After seeing the simulations in class on the computer, it seemed almost impossible.
Sheila: I know this because we saw it on the computer program in class.

Dixie: When we did over 5000 tests via the software program, we STILL didn't get the lower #.
Daisy: When we did the test on the computer it took 5000 (trials).
Loni: I remembered in class the computer simulation took 5000 (trials).
Sandy: I was thinking about the simulation in class and how many trials we had to enter in the computer.
Frida: I based it on the activities we have done in class with computer program as well as hands-on activities.

More than a few sample responses have been shared to emphasize the impressions that the class experience made on the students.

4.3. INTERPRETATIONS OF VARIATION

In addition to reasoning about expectations, subjects also revealed changes in how they thought about their interpretations of variation according to causes, effects, and influences on expectation and variation.

Causes For causes of variation, while there was heavier attention paid by some students in the PreSurvey and PreInterview to the physical nature of performing the flips or spins in the probability context (or drawing the samples in the sampling context), there was relatively less concern with these human causes of variation in the PostSurvey and PostInterview. Prior to the class interventions, many subjects expressed concerns about how samples were drawn, coins were flipped, and spinners were spun. Particularly in the case of spinners, the class as a whole seemed initially skeptical about whether or not spinners could actually be a true random device. Some initial responses from the PreSurvey were about the use of two half-black and half-white spinners, and whether or not the chance of both spinners landing on black was 50%:

Molly: Only if the spinner starts spinning in between both is it a 50-50. I think.
Rosie: A lot I think depends on how you spin.
Sarah: I think it depends somewhat on where the spinner is started from and the spinner is not on the same point in both pictures.
James: Depends on the force used to spin, the resistance of the spinner, the direction of the spin.

The representative responses shared above help to illustrate the concern with how the user operates the spinner, hinting that the user can cause more or less variation depending on the technique used. James helped further explain his concern about spinners in the PreInterview:

James: I want to look at the engineering of the spinner, where do you start the spin, you know, I mean…. Do you start it in white, you know, the velocity, or the force… None of that really matters, I guess…
I: I'm asking…
James: I mean, it CAN matter of course, yeah. Well, of course, it WOULD matter, you know, I mean, you play like a game that has a spinner, and, if you're a kid, you know if you hit it just the right way, and you start it at just the right the spot, there's a chance of it being in one spot are greater than in another spot.

I: So this is very well-oiled spinner…Very, very fair spinner
James: Ok, so this is a GOOD spinner. Yeah. Ok. A fair spinner. And the spinner is flat? A flat plane? It's a fairly spun game?

Rather than being contentious, James was expressing notions about fairness that were shared by others in class. Once the class interventions got underway, it became apparent that a major point of discussion was how children (and themselves) might strive to impede variation by, for example, flipping a coin in a certain way, or hitting the needle of a spinner. Even in sampling candies from a jar, subjects wondered about the plausibility of reaching into the jar in a special way so as to minimize variation. After the interventions, in the PostSurveys and PostInterviews, very little was expressed by the subjects about their concerns over causes of variation. One reason for the lack of commentary may be because the class had seemed to resolve the issue of deliberate causes. That is, they clearly knew a great deal about how children might tamper or try to tamper with random devices, and even in their own activities the subjects sometimes struggled with one another over how to fairly use the devices

The class seemed to come to a consensus that the point of doing a probability experiment really hinged on the assumption of randomness, and that their job was to help and not hinder the natural variation of outcomes. That is, they were not to try and spin a certain way to get a certain result, they were just supposed to spin and let the pointer land where it may. Thus, there may have just developed an acceptance of the myriad forms of physical, deliberate causes of variation. Having expressed their concerns in the PreSurvey and PreInterview, and having discussed these concerns in the class activities, they may have reconciled the issue of physical causes, leaving them more sensitive to the natural random variation inherent in the probability activities.

Effects As a part of the framework discussed earlier, the effects of variation were seen in terms of how students perceived probability situations and how they decided on their predictions. The focus of the *effects* component of the framework is therefore aimed at the effects on how students think, and a noticeable change reflected in class responses was a shift away from an "Anything can happen" and an "I don't know" mentality. In terms of trajectory of thinking, a precursor to how "Anything can happen" seemed to be the idea of how reality was different from theory. For example, in the PreSurveys and PreInterviews some students expressed the "Reality versus Theory" mindset in explaining their reasoning:

Daisy: Because probably outcomes aren't for sure outcomes.
Ross: Reality does not obey the estimates of probability.
Sergio: You are dealing with chance, like gambling. In theory there is probably an answer…But if you do it for real, 100 times, the numbers change but the ratios do not.

In the PreInterview, Ross was able to expand on his thinking, and he described a "probability-dictated reality, as distinct from described likelihoods." When asked for further explanation, he said: "I thought, okay, reality is going to impinge on the strict likelihood by a given thing." Thus, an effect of variation for Ross and others is that reality does not always match with what probability says should happen.

From discussions in class, it became apparent the "Reality versus Theory" mindset was held by many. However, a potentially unhelpful result of the "Reality versus Theory" mindset seemed to be that if theoretical predictions couldn't be counted on in reality, then

"Anything could happen" For example, in considering the prediction of probability outcomes in the PreInterview, some subjects were deliberating about what outcomes to choose:

> Sarah: Just choose randomly – Anything is possible.
> James: Well, they're all likely.
> George: You could just get any number.
> Sandy: Logically that's what my brain is telling me, is it can be absolutely anything.

A major problem with the "Anything can happen" mindset is that subjects who held this view tended to think of *all* outcomes not only as possible, but also as somewhat equiprobable. As Sandy said later on in the PreInterview, "I feel like it really can be anything. And so making a guess is just like… Just saying anything." Sandy's comment gives no regard to the relative likelihoods of different outcomes, and implies a complete lack of guidance in making predictions.

Along with the "Anything can happen" view, a strong undercurrent of the class discussions prior to the interventions swelled toward the idea that it wasn't possible to even make a prediction, which was likened to guessing – the "I don't know" mindset. The following excerpts illustrate what subjects wrote when asked to make predictions on the PreSurvey:

> Alice: You can make a prediction, but not a concrete answer.
> Leila: Always getting (25 heads) is hard to predict.
> Carrie: Hard to say. The odds are never exact.
> Frida: Couldn't hazard a guess, or could but it would be random.
> Rosie: This one I don't know. I have to do it physically.

The key feature that emerged from PreSurvey and PreInterview responses as well as from the class discussions prior to the interventions was that many students were extremely reluctant to make predictions, often using language to the effect that they "couldn't guess." The PreInterviews helped show that what was meant by the "I don't know" mindset was not really that students couldn't guess or predict, but that they couldn't know ahead of time whether or not their predictions would be correct:

> Emma: You just never know what you're going to get.
> James: It's impossible to know. Because you can't predict the future. I mean, I don't know what I'm going to get.
> Sandy: I dunno, I can't guess. I have trouble making guesses because… I can never know.

Sandy in particular encapsulated the view of many in the class, saying in the PreInterview that "you can never really guess. Because there's always a chance that any of those numbers could be anything." The trajectory of thinking held by many students prior to the class interventions was that (1) Probability theory may suggest a given result, but in reality results will vary; (2) Since results can vary, anything is possible, even to the point of being equally likely; (3) Since anything can happen, one can't know ahead of time what will occur, so it isn't possible to know ahead of time what will occur.

Thus, it is the effect of variation upon perceptions ("Anything can happen") that also interferes with the effect of variation upon decisions ("I don't know"). In other words, it

is the variation inherent in the probability situations that results in uncertainty, leading in turn to the difficulty students have with making a prediction. The hypothesis is that variation (and the resultant uncertainty) means one doesn't know for sure what will occur, and if one doesn't know what will occur, then results could be anything, thus confounding expectations. What is really striking is that virtually none of the "Anything can happen" and "I don't know" views were expressed after the class interventions. The discussions from the class, along with PostSurvey and PostInterview responses, suggest that most subjects thought that although one may not know for sure about a given outcome, one can still make reasonable statements of expectation. Also, subjects had less difficulty in making choices and decisions in the PostInterviews, and choices were more reasonable than in the PreInterviews.

Influencing Expectations and Variation Finally, the conceptual framework dimension of influencing expectations and dimensions came through more strongly and credibly after the class interventions, chiefly in the way that subjects referred to the number of sets of spins used in the PostSurvey and PostInterview. The number of sets was related to the average, extremes, and the overall distribution of results from multiple trials (sets of spins), with richer notions being expressed after the class interventions. Prior to the class interventions, for example, Sandy mentioned expecting "an average of 25, if you did many of these sets," and Ross agreed that "if you're going to see a range of results, the average of that range will be 25, but not every result will be 25." The ideas put forth by Sandy and Ross were shared by others in the class who thought that even if results varied, the average for multiple sets would or should be the expected value (or close to that value). After the class interventions, there were more comments that reflected how performing more sets of spins would draw the cumulative average closer to the expected value:

> Sandy: The more that you would do these sets of 50 spins, the more it would probably come back towards that 25.
> Sergio: The more times, the closer it will be toward 25.
> Loni: The more times he spins, the closer he will actually get to the 50/50 chance.
> Sheila: The more he spins the closer the results will match the probability (1/2).
> Maya: It will be even closer to 25 because of the Law of Large Numbers.
> James: So, the theoretical should come close to the experimental… Over the long run, if we do enough trials, chance are, they'll come pretty close if we do a fair number of sets.

The richness of the type of thinking really comes through in James' comment above, as it exemplifies the ideas from the class discussions how experimental probability relates to theoretical. Most importantly, instead of thinking that the average needed to always be the expected value, after the class interventions there were more comments such as George's: "Your mean and median will probably get closer and closer – the more and more you do – you know, the closer and closer you would get to 25." In particular, George pointed out in the PostInterview, with fewer numbers of sets "you're going to have a lot more variation in where the median and the mean are going to go." George's remarks, shared by others in the class, clearly show an appreciation for how even averages can vary.

As for influencing the extremes with performing more sets, there was an appreciation both before and after the interventions of how more sets would expand the range, but the

comments were rather thinly expressed at the outset of the course. Typical notions were that "the range will increase with increasing attempts" (Cammy), and "the more sets you do, the more often you'd expect to get that low chance" (Sandy). After the interventions, again the same kind of idea was expressed about an increasing range happening with more and more sets, but the language was more sophisticated:

Ross: As the number of trials goes up, so expands the range of possible outcomes towards the extremes.
Sandy: You would expect with more sets you do, the more sort of outliers you would get, or the 'unexpecteds' you would get.
Emma: The more sets, the more opportunity you have for outliers.
George: The more you do, the better chance of getting those extreme numbers.
Rocky: And so, the more sets you do, the more opportunity that exceptional event has of occurring, the more chance there is of getting an outlier, or an extreme value.

Of course, the language of outliers and extremes was a part of the course discussions, and the class had seen how doing more sets had in fact made it more likely to get an outlier, so it made sense to see these ideas expressed after the interventions.

Regarding distributional thinking, mostly primitive notions came through prior to the interventions, with some students mentioning how results might vary with more sets. However, the few responses were not very specific:

Daisy: The more you do something, the more chances you have that it's going to vary from the percentage.
Dixie: If you have more friends doing it, then I think there's more chance of more variation.
Jackie: The more sets done, the more likely you will get less likely results.
Julie: The more people that do the experiment, the more varied the results.

After the interventions, responses were more articulate. For example, Daisy expressed that "the more times you do it, you'll have variations on each end, which might get wider, but you'll have more in the center, around the 25." Julie mentioned that with more sets, "the results would get tighter, the grouping would accumulate around 25." More importantly, subjects gave reasonable comments aimed at the shape of the underlying distribution:

Daisy: The more pulls you do, the more evenly shaped your graph is going to be. Where fewer pulls, you're going to have a little more unevenness in your curve.
Ross: The more trials run, the more normal the distribution, but the chance of outliers also increases. I expect to see a certain bell curve, given more trials.

Note how Daisy mentions both the influence of more and of fewer sets, with fewer sets attributed to an "unevenness" in the graph of results. Others in class agreed with this idea, and Sandy expressed that with fewer sets, "you expect there to sort of be this more random look to it, so it's going to look a little bit more scattered." With more sets, Sandy thought "it would become more conformed to this perfect bell-curve, and that it would pull out a little bit more" meaning the tails of the graph would extend further. Even

though Ross and Sandy appropriately discuss bell curves in the context of the probability problems, they and others in class also tended to use the language of a "bell curve" or a "symmetric" distribution even on other sampling and probability questions where the underlying distribution was not normal.

5. SUMMARY AND IMPLICATIONS

In this final section, first a summary of the main insights of this study is given, and then implications for future teacher training are discussed

5.1. MAIN INSIGHTS

One of the main insights from this research is how subjects became more attentive to variation throughout the course. Written class responses showed improvements from the PreSurvey to PostSurvey, when evaluated according to rubrics that placed higher value on responses recognizing variation in probability situations. Each of the subquestions of *One Set, Compare Sets,* and *Six Sets* showed overall class improvements regarding what subjects expected and why (reasons for expectations). In *One Set*, more subjects gave range expectations and gave evidence of reasoning using both variation and proportions. For *Compare Sets,* more subjects incorporated a range or some kind of spread into their explanation of why results would not necessarily be the same for the results of the second set as on the first set. With *Six Sets,* students gave more appropriate choices that were backed up with reasoning explicitly using proportions and variation. Despite the improvements shown by subjects towards the end of the course, it should be acknowledged that there were still areas in which a substantial percentage of students showed a less than optimal performance. For example, as Table 3 showed earlier, on *One Set* almost half of the students still gave 25 as an answer on the PostSurvey, rather than some other response that might better acknowledge the effect of variation.

Another main insight from this research is the usefulness of the conceptual framework in characterizing the thinking of elementary pre-service teachers. While the coding rubrics were useful for gaining a quantitative picture of overall class changes, the evolving framework was a useful lens for looking more closely at key aspects of reasoning about variation which changed over the course. The interview responses combined with more detailed survey responses helped paint a more detailed picture of the richer understanding that emerged from subjects in terms of their expectations and interpretations of variation. Overall, subjects drew from their collective learning to better reason in terms of what they expected and *why*. Their predictions in the PostSurveys had better attention to range considerations and less emphasis on repeated values for results, particularly involving the expected value. Their reasoning included appropriate balancing of proportional thinking along with an appreciation of variation in expressing what was likely or probable. Class experience clearly had an influence on the reasoning of many students after the interventions, particularly the use of computers. Their interpretations included a reconciliation of physical causes of variation, leading them to focus more on natural causes of variation, namely the randomness inherent in the probability situations. Instead of interpreting variation as simply leading to an "Anything can happen" mindset, accompanied by an "I don't know" regard for making predictions, more students were able to express reasonable predictions. Also, students showed some reasonable interpretations of the effect that performing more trials might have on the cumulative average, presence of outliers, and shape of distribution of results.

5.2. FUTURE IMPLICATIONS

Implications for teaching elementary pre-service teachers include the suggestion that having hands-on activities, bolstered by small-group and whole-class discussion focused specifically on variation, can be a powerful way to move them toward a better appreciation of how variation plays a role in statistical thinking. The class interventions involved all three main aspects of understanding variation (expecting, displaying, and interpreting) in the contexts of sampling, data and graphs, and probability situations, all of which are important for elementary school children to address. If school teachers are to shape their lessons so as to encourage statistical thinking in their own students, then university teacher training programs need to provide an environment where pre-service teachers can learn in a similar way that they themselves will aim to teach (National Council of Teachers of Mathematics, 1991). In the environment where this research took place, the teaching philosophy of the course encouraged a great deal of discourse among students, which served to naturally provide springboards from which class discussions of variation could emerge. Also, a key design component of the surveys, interviews, and class interventions was that subjects were expected to make conjectures and discuss their reasoning before actually doing any activities. By laying out ahead of time what everyone in class thinks, groundwork can be established for making comparisons after actual data have been collected by doing the probability experiments. The computer simulations, brought out only after students have physically run simulations themselves, also seem to hold much promise for getting subjects to understand long term trends. Elementary pre-service teachers, like the children they will one day teach, need to investigate variation in probability settings by conjecturing, reasoning together, doing experiments, and discussing findings. The task for teacher educators is to continue to develop ways to structure their college classes to support elementary pre-service teachers' reasoning about variation.

However, in designing instruction for elementary pre-service teachers, it is important to keep learning more about the initial conceptions of variation that they hold, and how those conceptions change with different instructional interventions. That is, there is an iterative sense in the way instruction for pre-service teachers is designed and then refined based upon what has been learned about how they think about variation. The research described in this paper has been largely exploratory because little has been known about the conceptions of variation held by elementary pre-service teachers. An implication for research is that more needs to be learned about how elementary pre-service teachers' conceptions of variation compare with those of elementary students. For example, what are some similarities and differences in the responses of elementary pre-service teachers and school children? How can elementary pre-service teachers increase their own knowledge of variation while also learning how children reason about variation?

6. CONCLUSION

As research in the field of statistics education advances, one goal is that teacher education can improve not only the subject matter knowledge of elementary pre-service teachers, but also the pedagogical content knowledge of teaching about variation. Steps toward improved pedagogical content knowledge can certainly be informed by recent research about how pre-college students learn. Meanwhile, steps toward improved subject matter knowledge can be informed by a consideration of the conceptions of variation held by pre-service teachers as they enter university programs. Collective discourse in the class, bolstered by activities and simulations targeted at eliciting conceptions of variation

and developing these concepts, hold promise as a way of building elementary pre-service teachers' knowledge while also reflecting the kinds of practice they themselves will want to demonstrate in their own classrooms.

REFERENCES

Canada, D. (2004). *Preservice teachers' understanding of variation.* Unpublished doctoral dissertation, Portland State University, Portland, Oregon (USA). [Online: www.stat.auckland ac.nz/~iase/publications/dissertations/dissertations.php]

Cobb, G., & Moore, D. (1997). Mathematics, statistics, and teaching. *American Mathematics Monthly, 104*(9), 801-824.

Finzer, W. (2001). *Fathom™ Dynamic Statistics* [Computer software, v. 1.1]. Emeryville, CA: KCP Technologies.

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92-99. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_Garfield_BenZvi.pdf]

Hammerman, J., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41. [Online: www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Hammerman_Rubin.pdf]

Jolliffe, F., & Gal, I. (Eds.). (2004). *Statistics Education Research Journal, 3*(2). [Online: www.stat.auckland.ac.nz/~iase/serj/SERJ3(2).pdf]

Jones, G., Mooney, E., Langrall, C., & Thornton, C. (2002). Students' individual and collective statistical thinking. In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society, Cape Town, South Africa.* [CD-ROM] Voorburg, The Netherlands: International Statistical Institute.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-451.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59-98.

Konold, C., & Miller, C. (1994). *ProbSim: A Probability Simulation Program.* Santa Barbara, CA: Intellimation Library for the Macintosh.

Makar, K., & Canada, D. (2005). Pre-service teachers' conceptions of variation. In the *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education*. Melbourne, Australia.

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

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society, Cape Town, South Africa* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

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

Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal, 3*(2), 84-105. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Reading.pdf]

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201-227). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Shaughnessy, J. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Bidduch & K. Carr (Eds.), *Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 6-22). Rotorua, New Zealand: MERGA.

Shaughnessy, M., & Arcidiacono, M. (1993). *Visual encounters with chance (Unit VIII, Math and the mind's eye)*. Salem, OR: The Math Learning Center.

Shaughnessy, J.M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society, Cape Town, South Africa* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy, J. M., Ciancetta, M. & Canada, D. (2004). Types of student reasoning on sampling tasks. In the *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*. Bergen, Norway.

Shaughnessy, J., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgement of statistical variation. Presentation in *There's More to Life than Centers*. Presession Research Symposium, C. Maher (Chair), 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4-14.

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

Truran, J. (1994). Children's intuitive understanding of variance. In J. Garfield (Ed.), *Research Papers from the 4th International Conference on Teaching Statistics (ICOTS 4)*. Minneapolis, MN: International Study Group for Research on Learning Probability and Statistics.

Watson, J., Kelly, B., Callingham, R., & Shaughnessy, J. (2002). The measurement of school students' understanding of statistical variation. *The International Journal of Mathematical Education in Science and Technology*, *34*, 1-29.

Watson, J., & Moritz, J. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145-168.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67,* 233-265.

DANIEL L. CANADA
Eastern Washington University
203 Kingston Hall
Cheney, WA 99004
USA

ASSOCIATION OF COURSE PERFORMANCE WITH STUDENT BELIEFS: AN ANALYSIS BY GENDER AND INSTRUCTIONAL SOFTWARE ENVIRONMENT

J. RICHARD ALLDREDGE
Washington State University
alldredg@wsu.edu

GARY R. BROWN
Washington State University
browng@wsu.edu

ABSTRACT

The effect of educational technologies on learning is an area of active interest. We conducted an experiment to compare the impact of instructional software on student performance. We hypothesize that some of the impact on student performance may reflect the influence of the technology on student subject-related beliefs and that those beliefs may differ by gender. We desired to assess how course performance may be associated with student beliefs, and how the association may differ depending on instructional software environment and gender.

Keywords: *Statistics education research; Instructional software; Student beliefs; Gender*

1. INTRODUCTION

1.1. BACKGROUND

An experiment in an algebra-based introductory statistical methods course presented an opportunity to assess the influence of an instructional software environment on the association between student beliefs and subsequent course performance. The influence of student gender on the connections between belief, performance and software environment is also of interest. The motivation for this investigation is stated by Gal, Ginsburg, and Schau (1997), "Lastly, in order to make the learning of statistics less frustrating, less fearful, and more effective, especially among college students but also at earlier stages, further attention by statistics educators should be focused on the attitudes and beliefs students bring into statistics education experiences, how they develop and change during their educational experiences, and the impact they have on students' achievement, persistence, and eventual application of their new knowledge and skills." In this study, beliefs about quantitative confidence, general academic confidence, quantitative background, and the importance of quantitative skill to future success were measured with a pre-course self-assessment (Appendix). Here beliefs are defined as individually held ideas about statistics, about oneself as a learner of statistics, and about the social context of learning statistics (Gal et al., 1997). Among the questions of interest are: 1) is there an association between pre-course beliefs and course performance? 2) does evidence of association remain stable throughout the course? 3) does the association

differ for females and males? 4) does the association depend on the instructional software package used? The answers to these questions have implications for designing intervention strategies for improving the teaching and learning of statistics.

1.2. PREVIOUS WORK

Research investigating student beliefs about science, mathematics, and statistics has been conducted by a number of authors (Gal & Ginsburg, 1994; Shamos, 1995; Seymour & Hewitt, 1997; Wisenbaker & Scott, 1998). Much of this work suggests that capable students are overtly discouraged from their interest or potential interests in science and mathematics. Negative beliefs can impede learning, hinder development of useful intuitions, and reduce application outside the classroom (Gal & Ginsburg, 1994). Most theories on academic motivation involve the premise that lack of self-confidence leads to a reluctance to try (Cross & Steadman, 1996). Rouse (1995) notes many negative beliefs among students about mathematics, including a lack of confidence in their ability to do mathematics. Also students' understanding, retention, and application of what is taught, and their motivation to learn, depends upon their sense of why this subject is necessary or useful. Moore (1997) states that the first topic within the course must be motivational for students, that is, an explanation of why students need to understand the material. Although distinctions can be made about the influence of negative beliefs in science education, mathematics education and statistics education, Gal et al. (1997) note that beliefs, achievement, and persistence influence each other in statistics education in ways similar to mathematics and other areas. There are differences as well. Huang and Brainard (2001) found female students' self-determinants of mathematics self-confidence to be different from factors that determine science self-confidence. Sax (1994) notes that traditional predictions of mathematics confidence operate differently for males and females and for science and nonscience fields at college entrance. Clark (1994) examined the effect of context on performance, for example, the teaching of statistics to first year university males who have a nonphysical sciences interest.

In addition to potential differences in the association between beliefs and course performance due to gender differences, field of interest differences, or science, mathematics, and statistics focus, there may be differences due to instructional materials employed. Shaughnessy (1992) suggested using computer software to change student beliefs. Moore (1997) proposes that video may be used to change the beliefs of viewers at a subconscious level so instructional software that includes carefully constructed video components may be more effective at changing beliefs than software without video clips. Harwood and McMahon (1997) concluded that integrated video media curriculum intervention can positively affect achievement and attitudes among high school chemistry students. Nevertheless, as Forbes (1996) notes, it is unlikely that any one technique will suit all learners. Adaptive technologies are frequently cited as an important way to address this challenge and others associated with improving instruction (National Science Foundation, 1996; Derry, 1992; McCalla, 1992). In addition to presentation of learning content in these technologies, much attention has been placed on the importance of the design, particularly focused on the user interface and ease of use (Nielsen, 2000; Reigeluth, 1999; Shneiderman, 1998; Ware, 2000). Finally, in contrast to the clamor associated with the arrival of technologies in education Zemsky and Massy (2004) counter with sobering evidence and argument that technology does little to revolutionize education, further suggesting that pedagogy and implementation are the only salient variables.

Based on this previous work we hypothesized that different instructional software environments that reflect different pedagogies would influence the association between student beliefs and student course performance in different ways. Furthermore, we hypothesized that the impact of the instructional software environment on the association would be different for females and males.

2. MATERIALS AND METHODS

2.1. THE EXPERIMENT

The experiment was implemented in an introductory algebra-based statistical methods course. This course satisfied a general education requirement for mathematics proficiency at Washington State University and satisfied a requirement of many departments. Students could enroll in the course with either a Math or Stat prefix, depending on their department's policy. Course content included material concerning methods for producing data, summarizing data graphically and numerically, describing and quantifying relationships between variables, measuring uncertainty with probability, sampling distributions, confidence interval estimation and hypothesis testing for proportions and means, and analysis of count data. The students came from broad backgrounds of previous mathematical and statistical knowledge and current academic interests. Two-thirds of the 172 students were female and 95 percent were between the ages of 18 and 24. One-third of the students had undeclared majors so student interest as evidenced by major was not included in analyses.

The course consisted of three hours of lecture instruction per week and a two-hour weekly laboratory session. There were two lecture sections of the course. One section was divided into three laboratory sections and the other larger lecture class was divided into six laboratory sections. Each laboratory section was assigned one of two instructional software packages to be used in the laboratory for the entire semester. To reduce instructor influence on overall differences among the beliefs and performances of students, a single instructor volunteered to teach both lecture sections of the course. The same textbook was used for both lecture sections. All three teaching assistants were assigned two laboratory sections from one instructional software package and one laboratory section for the other software package. All students in a laboratory section used the same instructional package. Students individually selected a lecture section and a laboratory section associated with that lecture section prior to the beginning of the term. For administrative convenience the three laboratory sections associated with the smaller lecture class used one package and the six sections associated with the larger lecture class used the other. Therefore there is a potential confounding effect of lecture and package even though the same instructor taught both lecture sections. No students switched lecture sections and hence software package associated with different laboratory sections during the term. Because the treatments were applied to laboratory sections, rather than to individual students, the nine laboratory sections were considered the experimental units for comparing instructional packages.

Two instructional software packages, ActivStats® and CyberStats™, were the treatments used in this study. ActivStats presents an introductory statistics course by integrating video, simulation, animation, narration, text, interactive experiments, and a statistics package into a learning environment (Addison Wesley Interactive, 1998). Product information accompanying ActivStats claims that students will experience real world examples, learn key statistics concepts through specially designed simulations, and practice with interactive experiments. CyberStats is a Web based textbook for an

introductory statistical methods course that features learning through interaction (CyberGnostics, 2004). Students interact with simulation and calculation applets. On the CyberStats web page the following principles are listed: learning by activity and discovery, real data in real-world settings, and a stress on conceptual understanding. Each package contains its own version of a computational statistics program that both interfaces with the topical lessons, and is available for use independently of the instructional activities. CyberStats is a world-wide-web based program. Students pay a fee for a password that gave them access to the material CyberStats for the duration of the academic term while students in the other treatment group purchased ActivStats on a CD-ROM. The cost for each package was approximately the same. These packages were chosen because we agree with Lee (1998) that introductory statistics should be taught using real world data, student activities, and computer technology. The decision not to use a formal control group with no instructional software treatment is consistent with an approach that assumes that there will be impacts and they will be different for the different instructional methods.

Despite similarities in the two software packages they reflect two distinct instructional strategies. ActivStats embodies design principles that reflect assumptions that learners benefit from a greater contextualization of the problems, a contextualization that situates the learning of statistics in word problems, and it places a conspicuous emphasis on organizing the learning of statistics around the primacy of broad concepts. The interface, consistent with those assumptions, provides links to videos that explore the context in which the statistical analysis will be provided, and the statistics are organized around concepts like "understanding data, understanding relationships, and generating data." For instance, instead of introducing the concept of regression, the organization subordinates the statistical methods to the umbrella concepts of relationships between things, and it presents videos. For example, a short video on the plight of the manatee is used to introduce the relationship of the animal to human incursions in the Everglades. In this context, regression is introduced as a tool to examine the relationships between human incursion and a declining animal population.

The CyberStats package reflects principles that hold the importance of the mathematical underpinnings of statistics. The different statistical methods shape the organization of the material, moving from the more basic principles to the more complex. The interface is designed to present the information about the statistical concept sequentially, including definition of terms. It then presents opportunities to practice the procedure. In addition, the package integrates the mathematical and statistical concepts with interactive models that demonstrate the graphical representation of the concept.

Students were directed to use selected material from their assigned software package during each laboratory session. They were also instructed to do selected laboratory homework exercises from their assigned package. The laboratory homework exercises counted toward their course grade. The selected material related to lecture topics presented during the class meetings prior to the scheduled laboratory.

2.2. INSTRUMENTS

At the first laboratory session, the students completed a questionnaire with 39 questions addressing issues of quantitative, verbal, and academic confidence. The questionnaire also addressed computer proficiency and students' feelings considering applications of statistics and general academic study to their future. The survey was modeled, with permission, after the Teaching Goals Inventory (Angelo & Cross, 1993). Angelo and Cross drew in particular on work by Kulik (1976) and Bowen (1977) to shape

their work on students' reactions to instructions. The aspects of the Teaching Goals Inventory that focused on attribution of responsibility for learning were particularly useful in our adaptation of the instrument. In addition to extracting and adapting questions from the Teaching Goals Inventory, we focused questions specifically on issues of general confidence and beliefs toward learning and toward confidence in mathematics and statistics in particular. We focused several questions about students' confidence, in order to explore issues that research suggests are promising for improving student performance, though there are also indicators in that work that improving confidence alone may not improve student performance. (Leder, Pehkonen, & Törner, 2002).

The pre-course questionnaire is presented in the Appendix. A similar questionnaire was given during the final laboratory session. In addition to questions about confidence and future applications of statistics the post-course questionnaire asked students to evaluate the instructional package they used during the course. An analysis of the difference between post and pre-class responses due to educational software treatment may be found in Alldredge and Som (2002).

Assessment of student learning included two mid-semester tests and a final examination based on topics covered in both the lecture and laboratory portions of the class. Mid-semester tests consisted of short answer and multiple-choice questions and were administered in lectures. Students in one lecture section had 50 minutes to complete the tests while students in the other lecture section had 75 minutes to complete a longer test. Several questions asked students to comment on or explain their results in words. Students were allowed use of calculators, statistical tables, and one sheet of self-prepared notes. The take-home final test consisted of story problems where computer assisted calculations were necessary, as well as short answer and multiple-choice questions. The take-home final test was untimed, open book, and unsupervised. Students were instructed to work independently and had one week to complete the final test. An additional assessment of student learning used total course points including all tests, final test, scores compiled from in-class and laboratory activities, lecture and laboratory homework assignments, and two class projects. The projects, although containing statistical analysis, were largely written works and graded for pertinent statistical content and quality of writing. Course grade, based on total points, was also used in analyses. Students' pre-course quantitative and verbal skills were assessed through SAT (formerly known as Scholastic Aptitude Test) verbal score, SAT mathematics score, and SAT total score. The SAT is a three-hour test that measures verbal and mathematical reasoning skills that is administered to secondary school students. Many colleges and universities use the SAT as one indicator of a student's readiness to do college-level work (SAT I, CollegeBoard.com).

2.3.STATISTICAL METHODS

In order to reduce the dimensionality of the questionnaire and identify the underlying patterns of variation in the data set, a multivariate principal component analysis (PCA) was conducted. A mixed model analysis of variance was used to explore the association between course performance and student pre-course beliefs. Specifically, analysis of variance and covariance were used to test for association between factor scores identified by the PCA and course performance, while considering the effect of instructional package used and gender. SAT mathematics, SAT verbal, or SAT total scores were used as covariates in the mixed model analysis of variance. Spearman correlation coefficients were computed for the ActivStats and CyberStats laboratory sections to measure the strength of the monotonic relationship between factor scores and course performance

(Hays, 1973, p. 787). We also tested the association between pre-course questionnaire item responses and overall course grade with the Jonckheere-Terpstra (JT) statistic (Hollander & Wolfe, 1973) for all laboratory sections combined and for the ActivStats and CyberStats laboratory sections separately. This statistic allowed testing a directional hypothesis between each item on the pre-course questionnaire and the final course grade. All analyses were completed for females and males separately to explore gender differences in the associations.

3. RESULTS

The principal component analyses produced a pattern and size of coefficients in the varimax rotated factor pattern that allowed three new variables that were linear combinations of the original response variables to be identified with labels. One of the linear combinations identified was composed of questionnaire items 3, 4, 6, 8, 12, 19, and 20 that are related to the student's self-reported concern about their ability to do mathematics (Mathematics Concern). A second factor, consisting of questionnaire items 1, 2, 17, was related to feelings of general confidence by students in their ability to do well in school (General Confidence). The third linear combination identified by the principal component analyses involved items 9, 11, 13, 15, 18, 22, and 24 that relate to past help and the applicability of mathematics, statistics and computer skills to their future careers (Math Commitment).

The presence of significant interactions in the mixed model analysis of variance indicated there were differences in the association between principal component factors and course performance measures depending on the educational treatment. Inclusion of SAT mathematics, SAT verbal, or SAT total scores as covariates in the mixed model analysis of variance usually markedly reduced the level of significance between a principal factor and the course performance score. This results from the highly significant correlation between SAT scores and the Mathematics Concern and General Confidence factors. Separate analyses for females and males showed different associations between principal component factors and some course performance measures prompting a separate consideration of the associations for each of the four instructional package-gender situations. That is, associations between pre-course beliefs and course performance are presented and compared for each of the groups: ActivStats-Female; ActivStats-Male; CyberStats-Female; CyberStats-Male. The number of students in each group having complete data for this analysis was 38, 21, 68, and 29, respectively.

Spearman correlation coefficients between the Mathematics Concern factor and test performance were negative throughout the course for all groups except the CyberStats-Male group (Table 1). That is, students who expressed more concern with their ability to do mathematics tended to have lower scores on all tests with the exception of the male CyberStats group. In the ActivStats group, this negative association was stronger for males than for females. The CyberStats-Female group had a significant negative association between all performance scores and Mathematics Concern while CyberStats-Male group had non-significant associations that were positive except for test 1. We note that Spearman's correlation coefficient does not provide information about independence of variables but rather is used here to provide a measure of association, namely, the direction and strength of the monotonic relationship between variables.

Correlation coefficients between General Confidence factor scores and test performance were significantly negative for the ActivStats-Female group (Table 1). In contrast the correlations were significantly positive for the ActivStats-Male group for all performance scores with the exception of the final test. In fact, the strength of the

association decreased throughout the course. Females in the CyberStats group had a significant positive relationship between General Confidence and test performance for tests 1 and 2 but the correlation decreased to 0.067 for the final test. Males in the CyberStats group had a non-significant association between test performance and General Confidence throughout the course.

The correlation coefficients for the Math Commitment factor with scores on all tests, as well as total course points, was generally negative, but not significant, for both males and females (Table 1).

The Jonckheere-Terpstra test revealed associations between several items on the pre-course questionnaire and final course grade. Some associations were consistent for both males and females for both instructional packages, while others indicated differences in significance depending on gender and instructional package. Table 2 shows the type of association found between final course grade and selected questionnaire items related to the three factors identified above. Notice that items 12 and 17 that relate to General Confidence had significant associations with course grade for ActivStats-Male group but not for the CyberStats-Male group (Table 2). For females, items 1 and 17 were significantly associated with final course grade for the CyberStats treatment group but not for the ActivStats group. Items 4, 6, and 19 that relate to Mathematics Concern all had significant associations with course grade for the CyberStats-Female but not for the ActivStats-Female group (Table 2). For males, item 4 had a significant association with final grade for the CyberStats group and a significant association for the ActivStats group. Like the females, males showed an association between item 6 and final grade for the CyberStats group. There were only a few items identified as being strongly related to Math Commitment that had a significant association with course grade (Table 2).

4. DISCUSSION

The findings related to gender, confidence, and preparation reported in the previous section suggest that there are complex associations between pre-course beliefs and course performance, but a more compelling finding is that associations between beliefs and course performance are not necessarily stable throughout the course. The implications of construct instability, therefore, underscore the complexity of the gender differences in the associations between pre-course beliefs and course performance. This complexity suggests that using technological interventions to mediate learning requires insights into how, and when, students absorb information provided by instructional technology.

The software packages—or the contexts of learning represented in the different designs of these two software packages—influence the associations between beliefs and learning outcomes differently. Specifically, ActivStats seems more effective in ameliorating the effect of Mathematics Concern on course performance compared to CyberStats for females. It may be that ActivStats, with its greater focus on contextualizing the presentation of statistical concepts, may better amend the negative attitudes females have towards mathematics than CyberStats. In contrast, for males it appears that ActivStats allows the negative association between Mathematics Concern and performance to persist while CyberStats may alter the negative association. Males may find that the more direct, linear format of CyberStats alleviates their concerns about their ability to do mathematics.

Table 1. Spearman correlation (p-values) between factor scores and course performance

ActivStats-Female			
Item	Mathematics Concern	General Confidence	Math Commitment
Test1	-0.333	-0.306	0.007
	(0.047)	(0.070)	(0.969)
Test2	-0.401	-0.321	-0088
	(0.015)	(0.057)	(0.609)
Final test	-0.170	-0.295	-0.038
	(0.323)	(0.081)	(0.827)
Total	-0.361	-0.320	-0.038
	(0.030)	(0.057)	(0.827)
Grade	-0.244	-0.353	-0.184
	(0.151)	(0.035)	(0.283)

ActivStats-Male			
Item	Mathematics Concern	General Confidence	Math Commitment
Test1	-0.544	0.504	-0.323
	(0.013)	(0.024)	(0.165)
Final	-0.406	0.304	-0.131
	(0.076)	(0.192)	(0.582)
Total	-0.496	0.436	-0.164
	(0.026)	(0.055)	(0.490)
Grade	-0.514	0.470	-0.068
	(0.020)	(0.037)	(0.775)

CyberStats-Female			
Item	Mathematics Concern	General Confidence	Math Commitment
Test1	-0.291	0.401	-0.033
	(0.023)	(0.001)	(0.799)
Test2	-0.352	0.246	-0.047
	(0.005)	(0.056)	(0.219)
Final test	-0.291	0.067	0.047
	(0.023)	(0.607)	(0.715)
Total	-0.398	0.239	-0.015
	(0.002)	(0.064)	(0.910)
Grade	-0.358	0.203	-0.035
	(0.005)	(0.117)	(0.788)

CyberStats-Male			
Item	Mathematics Concern	General Confidence	Math Commitment
Test1	-0.165	0.083	-0.235
	(0.410)	(0.681)	(0.238)
Test2	0.136	0.207	-0.071
	(0.499)	(0.299)	(0.726)
Final	0.170	-0.186	-0.283
test	(0.396)	(0.352)	(0.152)
Total	0.139	0.048	-0.206
	(0.489)	(0.814)	(0.303)
Grade	0.087	0.066	-0.357
	(0.665)	(0.744)	(0.068)

*Table 2. Direction and significance(*p≤0.10) of association between final course grade and questionnaire items*

	ActivStats		CyberStats	
Factor/Questionnaire Item	Female	Male	Female	Male
General Confidence				
1. I have confidence in my ability to do well on exams	+	+*	+*	+
2. I have confidence in my ability to write well	–*	+*	+	–*
17. When I apply myself, I do well in school	–*	+*	+*	+
Math Concern				
4. Math formulas confuse me	–	–*	–*	+*
6. My previous instruction in math was poor	–	–	–*	–
19. When I struggle with math I feel unintelligent	+	–	–*	+
Math Commitment				
11. I usually study math with friends	–	–	+	–*
13. My previous instructors are responsible for my attitude toward statistics	–*	+	–	–
15. Stat skills are essential to my future career	–	+*	–	–*
22. I spend a lot of time studying math	+	–*	+*	+
24. Computer skills are essential for my future success	–	+	–*	–*

The associations between General Confidence and course performance are different than those for Mathematics Concern and course performance. CyberStats appears to allow a stronger relationship between General Confidence and course performance than ActivStats for females while the opposite is true for males. For females the association is persistently negative for the ActivStats treatment group while for the CyberStats group the General Confidence relationship with course performance is positive, although it does decrease during the semester. For males ActivStats may encourage a more positive association between General Confidence and course performance than CyberStats. It may be that the focus on context in ActivStats provides connections to general feelings of confidence for males but not for females. For example, the links to videos that depict context, such as the plight of the manatee due to increased boat traffic, are used to introduce statistical methods of analysis. It may be that videos of real world problems, and effective statistical solutions, bolster confidence in males more so than in females.

The implications of the complex relationship between strategies that encourage confidence and those that improve performance emphasize a critical distinction: confidence and performance are not at all the same, especially for women. As Fennema (1996) notes, we do not know how confidence influences learning, but it has long been assumed that lower confidence contributes to gender differences in learning mathematics.

This raises serious issues about the efficacy of educational measures, tests, and instructional strategies that merit additional research. The complexity is underscored by the findings that although females had a significantly lower score on questionnaire item 3 (I have confidence in my ability to do math) they scored significantly higher than males on all exams except the first, and achieved more total course points. These results are consistent with other research (Rosser, 1989) indicating that even when females do well on exams they have a lower perception of their mathematics ability than do males. Further, it should be noted again that the combined scores for male and female students in laboratories that used ActivStats had significantly higher mean scores for all exams as well as total course points compared to students in the CyberStats laboratories (Alldredge & Som, 2002). Despite these generalizations, it appears that there are complex differences between males and females in the influence of differently designed software packages on the association between their beliefs and course performance. The technology that was designed to expand the context of statistics and that emphasized the methods of statistics through use of video components as ways to examine the context was more effective in terms of course performance for many students than was the technology that placed a more immediate focus on statistical methods, though the latter used examples as well. The distinction might be simplified. The more effective approach focused on statistics as a set of tools useful for examining the world; the less effective approach focused on statistics as an end, as content to be learned. However, the results obtained in this study indicate that in the realm of beliefs the effectiveness of the packages varied depending on student gender and dimension of belief.

What emerges is that persistent skepticism about the efficacy of technology as a way to improve learning is misdirected, and the findings in this study contribute to the growing body of research that argues that point. Researchers need to move beyond the simple question, “Does information technology work?” and examine instead the complex nuances of instructional design and the underlying strategies associated with that design, with or without technology. The differential findings in this study illuminate this point. Based on previous work (Zemsky & Massy, 2004) that identifies the salience of various pedagogical designs and implementation rather than the generic and more common tendency to lump all technologies into binary pronouncements, future research on intervention strategies to improve learning will benefit from attention to the complexity of the association between student beliefs and student achievement.

It is clear that more study is necessary. We have only part of the story here concerning how to change future practice. Perhaps combining the insights gained here with learning styles information would support recommendations about the future for use of instructional software.

ACKNOWLEDGEMENTS

We would like to thank the graduate teaching assistants as well as the students who participated in the experiment. In addition we would like to express our appreciation to the editor and referees for their insightful suggestions for improvement.

REFERENCES

Addison Wesley Interactive. (1998). ActivStats. Reading, MA: Addison Wesley [Online: www.aw-bc.com]

Alldredge, J.R., & Som, N.A. (2002). Comparison of multimedia educational materials used in an introductory statistical methods course. In B. Phillips (Ed.), *Proceedings of*

the Sixth International Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute.

Angelo, T., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers*. San Francisco: Jossey-Bass.

Bowen, H. R. (1977). *Investment in learning: The individual and social value of american higher education*. San Francisco: Jossey-Bass.

Clark, M. (1994). The effect of context on the teaching of statistics at first year university level. In L. Brunelli & G. Cicchitelli (Eds.), *IASE, Proceedings of the First Scientific Meeting* (pp. 105-113). Perugia, Italy: University of Perugia.

Cross, K. P., & Steadman, M. H. (1996). *Classroom research: Implementing the scholarship of teaching*, San Francisco: Jossey-Bass.

CyberGnostics. (2004). CyberStats. [Online: www.cyberk.com]

Derry, S. (1992). Metacognitive models of learning and instructional systems design. In M. Jones & P. Winne (Eds.) *Adaptive learning environments: Foundations and frontiers* (pp. 257-286). Berlin: Springer-Verlag.

Fennema, E. (1996). Mathematics, gender, and research. In G. Hanna (Ed.), *Towards gender equity in mathematics education* (pp. 9-26). Dordrecht: Kluwer Academic Publishers.

Forbes, S. D. (1996). Curriculum and assessment: hitting girls twice. In G. Hanna (Ed.), *Towards gender equity in mathematics education* (pp. 71-79). Dordrecht: Kluwer Academic Publishers.

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

Gal, I., Ginsburg, L., & Schau, C. (1997). Monitoring attitudes and beliefs in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 37-51). Amsterdam: IOS Press and the International Statistical Institute.

Harwood, W. S., & McMahon, M. M. (1997). Effects of integrated video media on student achievement and attitudes in high school chemistry. *Journal of Research in Science Teaching*, *34*, 617-631.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd edition). New York: Holt, Rinehart & Winston.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: John Wiley & Sons, Inc.

Huang, P. M., & Brainard, S. G. (2001). Identifying determinants of academic self confidence among science, math, engineering and technology students. *Journal of Women and Minorities in Science and Engineering*, *7*, 315-337.

Kulik, J. (1976). Student reactions to instruction: Memo to the faculty. Ann Arbor: University of Michigan.

Leder, G. C., Pehkonen, E. & Törner, G. (2002). *Beliefs: A hidden variable in mathematics education*. Hingham, MA: Kluwer.

Lee, C. (1998). An assessment of the PACE strategy for an introductory statistics course. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, & W-K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 1215-1222). Voorburg, The Netherlands: International Statistical Institute.

McCalla, G. (1992). The search for adaptability, flexibility, and individualization: Approaches to curriculum in intelligent tutoring systems. In M. Jones & P. Winne (Eds.), *Adaptive learning environments: Foundations and frontiers* (pp. 91-122). Berlin: Springer-Verlag.

Moore, D. S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review*, *65*, 123-165.

National Science Foundation. (1996). *Shaping the future: New expectations for undergraduate education in science, mathematics, and technology*. Washington, D.C: Government Printing Office.

Nielsen, J. (2000). *Designing web usability*. Indianapolis: New Riders Publishing.

Reigeluth, C. (Ed.) (1999). *Instructional-design theories and models: A new paradigm of instructional theory*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rosser, P. (1989). *The SAT gender gap: Identifying the causes*. Washington D.C.: Center for Women Policy Studies.

Rouse, L. P. (1995). Women and minorities in a social statistics course. *Journal of Women and Minorities in Science and Engineering*, *2*, 181-192.

SAT I, CollegeBoard (2004). [Online:http://www.collegeboard.com/student/testing/sat/about/SATI.html]

Sax, L. J. (1994). Predicting gender and major-field differences in mathematical self-concept during college. *Journal of Women and Minorities in Science and Engineering*, *1*, 291-307.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction*. Reading, MA: Addison Wesley.

Seymour, E., & Hewitt, N. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, CO: Westview Press.

Shamos, M. H. (1995). *The myth of scientific literacy*. New Brunswick: Rutgers Press.

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

Ware, C. (2000). *Information visualization: Perception for design*. San Francisco: Morgan Kaufman.

Wisenbaker, J., & Scott, J. S. (1998). A multicultural exploration of the interrelationships among attitudes about and achievement in introductory statistics. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, & W-K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 709-710). Voorburg, The Netherlands: International Statistical Institute.

Zemsky, R., & Massy, W. F. (2004). Thwarted innovation: What happened to eLearning and why. A Final Report for The Weatherstation Project of The Learning Alliance at the University of Pennsylvania in cooperation with the Thomson Corporation.

J. RICHARD ALLDREDGE
Department of Statistics
Washington State University
Pullman, WA 99164-3144
USA

APPENDIX: Pre-course questionnaire

Your feedback preceding this course will provide important and useful information for the course developers, the department, and the university. Please read the instructions carefully before giving your answers. Thank you for participating in this project.

Student ID #	Gender: M F
Your TA's name:	Your Major:
Year in School:	Minor (if applicable):

Part I: Background

Indicate how strongly you agree or disagree with each of the following statements:
(mark the appropriate circle, select only one response per question)

		Strongly agree	agree	somewhat	disagree	strongly disagree
1.	I have confidence in my ability to do well on exams.	□	□	□	□	□
2.	I have confidence in my ability to write well.	□	□	□	□	□
3.	I have confidence in my ability to do math.	□	□	□	□	□
4.	Math formulas confuse me.	□	□	□	□	□
5.	I have a good background in statistics.	□	□	□	□	□
6.	My previous instruction in math was poor.	□	□	□	□	□
7.	I am usually systematic in my approach to problem solving.	□	□	□	□	□
8.	I am usually well prepared for math exams.	□	□	□	□	□
9.	Math skills are essential to my academic success.	□	□	□	□	□
10.	I am generally good at visualizing concepts.	□	□	□	□	□
11.	I usually study math with friends.	□	□	□	□	□
12.	Math requires extensive mental discipline.	□	□	□	□	□
13.	My previous instructors are responsible for my attitude toward statistics.	□	□	□	□	□
14.	My family are pretty good in math.	□	□	□	□	□
15.	Stat skills are essential to my future career.	□	□	□	□	□
16.	People who are exceptionally good in math are often perceived as odd.	□	□	□	□	□
17.	When I apply myself, I do well in school.	□	□	□	□	□
18.	In the past, I have generally gotten help in math from family or friends.	□	□	□	□	□
19.	When I struggle with math I feel unintelligent.	□	□	□	□	□
20.	Most of my friends are better at math than I am.	□	□	□	□	□
21.	It is important to get to know students who are different from me in their cultural and socio-economic backgrounds.	□	□	□	□	□
22.	I spend a lot of time studying math.	□	□	□	□	□
23.	I am good in music.	□	□	□	□	□
24.	Computer skills are essential for my future success.	□	□	□	□	□

Part II: Technological Background

Rate your ability to do each of the following: (Circle the appropriate number from 1 – no knowledge/ability to 5 - expert user. Circle only one)						
		no knowledge/ ability				expert user
25.	send and receive voice mail	1	2	3	4	5
26.	create a word processed document on a computer	1	2	3	4	5
27.	program a VCR	1	2	3	4	5
28.	send and receive documents on a fax machine	1	2	3	4	5
29.	use a video camera	1	2	3	4	5
30.	use a spreadsheet or database program on a computer	1	2	3	4	5
31.	send and receive e-mail	1	2	3	4	5
32.	search for information on the Internet/World Wide Web	1	2	3	4	5
33.	program a computer using a programming language (such as Fortran, C, C++, or a database language such as Foxpro or Oracle, etc.)	1	2	3	4	5
34.	program a computer using a database language (such as Foxpro or Oracle, etc.)	1	2	3	4	5
35.	Create or edit a World Wide Web site (using such programs as html, java, etc.)	1	2	3	4	5
36.	electronically send and receive files by way of the computer (over a modem, the Internet/WWW etc.)	1	2	3	4	5

What type of computer do you use? (mark all that apply)

		Mac	Dos/ Windows	Windows/NT	Unix	Other (Please specify)	N/A
37.	at home?	☐	☐	☐	☐	☐ _____	☐
38.	at work?	☐	☐	☐	☐	☐ _____	☐
39.	in a university computer lab	☐	☐	☐	☐	☐ _____	☐

Part III: *In the space below, please answer the following question:*
What is the most important thing you hope to learn in this course?

PAST IASE CONFERENCES

1. SRTL-4 THE FOURTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING AND LITERACY
Auckland, New Zealand, July 2-7, 2005

We are pleased to announce the publication of *Reasoning about Distributions: A Collection of Current Research Studies*, a unique CD that contains research papers on reasoning about distributions presented at the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4) held in July 2005 at the University of Auckland, New Zealand. Many of these papers (all written in English) contain video segments (in English or with English subtitles) of student or teacher interviews, or learning interactions. The video segments and research studies provide a rich resource for researchers and teachers. Note: revised versions of some of these papers will appear in a special issue of the IASE *Statistics Education Research Journal* in November 2006. However, they will not contain video links and not all papers in the CD will be in this issue.

Due to ethical agreements with participants in the research studies, the video segments included on the CD are to be used only for research purposes. For any other purpose (e.g., professional development), permission must be obtained from individual authors. The CDs are available for a minimal cost that covers materials and postage. They may be purchased from Dr. Katie Makar (k.makar@uq.edu.au) at the University of Queensland (Australia).

OTHER PAST CONFERENCES

1. INTERNATIONAL CONFERENCE OF THE MATHEMATICS EDUCATION INTO THE 21ST CENTURY PROJECT: "REFORM, REVOLUTION AND PARADIGM SHIFTS IN MATHEMATICS EDUCATION" Johor Bharu, Malaysia, November 25 – December 1, 2005

The proceedings (available on the web, see below) contain the papers presented at the International Conference on Reform, Revolution and Paradigm Shifts in Mathematics Education held from November 25 to December 1st, 2005 at the Hotel Eden Garden, Johor Bahru, Malaysia. This Conference was organized jointly by Universiti Teknologi Malaysia and The Mathematics Education into the 21st Century Project - a non-commercial international educational project founded in 1986. Our Project is dedicated to the improvement of mathematics education worldwide through the publication and dissemination of innovative materials and ideas.

The title Reform, Revolution and Paradigm Shifts in Mathematics Education derives from Kuhn's seminal work on "The Structure of Scientific Revolutions" which made popular the concept of "paradigm shift," and the work of Lakatos which revealed the essentially creative and human nature of progress in science and mathematics. We hope these powerful ideas will help to inspire our conference. For further conference and project details email to Alan Rogerson (arogerson@inetia.pl).

Website with full paper access:
http://math.unipa.it/~grim/21_project/21_malasya_2005.htm

2. ASIAN TECHNOLOGY CONFERENCE IN MATHEMATICS, ATCM2005 Cheong-Ju, South Korea, December 12-16, 2005

The 10th Annual conference of Asian Technology Council in Mathematics (ATCM) on the theme Enriching Technology in Enhancing Mathematics for All was hosted by Korea National University of Education in Cheong-Ju, South Korea. The aim of this conference was to provide a forum for educators, researchers, teachers and experts in exchanging information regarding enriching technology to enhance mathematics learning, teaching and research at all levels. The conference covered a broad range of topics on the application and use of technology in Mathematics research and teaching.

Website with full paper access:
http://www.atcminc.com/mPublications/EP/EPATCM05/enter.shtml

FORTHCOMING IASE CONFERENCES

1. ICOTS-7: WORKING COOPERATIVELY IN STATISTICS EDUCATION
Salvador (Bahia), Brazil, July 2-7, 2006

The International Association for Statistical Education (IASE) and the International Statistical Institute (ISI) are organizing the Seventh International Conference on Teaching Statistics (ICOTS-7) which will be hosted by the Brazilian Statistical Association (ABE) in Salvador (Bahia), Brazil, July 2-7, 2006.

Information about ICOTS-7 is on the website: http://www.maths.otago.ac.nz/icots7 and in particular under the "Registration" tab, where there is information about registration, paper preparation, accommodation, and tours.

There are reduced fees for IASE/ISI members, and participants from Latin American or Developing Countries. There are 550 delegates registered from 60 countries. It promises to be an exciting conference. A large number of Brazilian school teachers are also attending a workshop just prior to the conference.

The abstracts for all the papers, plenary, invited and contributed, are on the website. There is also a list of all the posters. An Abstract Book with all this information will be provided in Salvador.

The conference timetable is also on the website and can be inspected to allow you to plan your activities during the conference. The timetable and other organizational details will be available in a Delegates' Handbook which will accompany the Abstract Book.

2. SRTL-5
THE FIFTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING AND LITERACY
The University of Warwick, Coventry, UK, August 11 - 17, 2007

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

The Forum's focus will be on informal ideas of inference rather than on formal methods of estimation and tests of significance. This topic is emerging from the presentations and discussions at SRTL-3 and 4 and is a topic of current interest to many researchers as well as teachers of statistics. As new courses and curricula are developed, a greater role for informal types of statistical inference is anticipated, introduced early, revisited often, and developed through use of simulation and technological tools. We encourage research papers that address reasoning about statistical inference at all levels of education including the professional development of elementary and secondary teachers.

2.1. TOPICS

We encourage submission of research papers that address questions such as the following:

1. What are the simplest forms of statistical inference that students can understand?
2. How does reasoning about statistical inference develop from the simplest forms (informal) to the more complex ones (formal)?
3. How can instructional tasks and technological tools be used to promote the understanding of statistical inference?
4. What are sequences of activities that can help student develop a conceptual understanding of statistical inference?
5. What types of misconceptions are found in students' reasoning about statistical inference?
6. What types of foundational knowledge and reasoning are needed for students to understand and reason about statistical inference?
7. How do students develop an understanding of the language used in describing statistical inference (e.g., significance, confidence)?
8. How does an understanding of statistical inference connect and effect understanding of other statistical concepts?
9. What are useful items and questions to use to assess understanding of statistical inference?

2.2. THE LOCAL SRTL-5 ORGANIZERS

Janet Ainley, janet.ainley@warwick.ac.uk
Dave Pratt, dave.pratt@warwick.ac.uk
For more information visit the SRTL-5 website: http://srtl.stat.auckland.ac.nz/

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

3.1. THEME

This satellite conference invites papers on all aspects of assessing student learning in statistics. For example, we expect to have papers on writing effective exam questions, on exam implementation strategies, and on alternative assessment methods such as projects, lab assignments, and writing assignments. We also encourage submissions on how to use assessment to improve student learning, and on developing and administering assessments items to conduct research into student learning. Proceedings will be available free at the publication page of IASE

3.2. CONFERENCE COMMITTEE

Brian Phillips (Australia) (Joint Chair and Joint Chief Editor) bphillips@swin.edu.au
Beth Chance (USA) (Joint Chair) bchance@calpoly.edu
Allan Rossman (USA) arossman@calpoly.edu
Ginger Rowell (USA) rowell@mtsu.edu

Gilberte Schuyten (Belgium) gilberte.schuyten@UGent.be
Larry Weldon (Canada) (Joint Chief Editor) weldon@sfu.ca
Local Organiser: Bruno C. de Sousa (Portugal) bruno@mct.uminho.pt
For more information visit the website at
http://www.stat.auckland.ac.nz/~iase/conferences.php?show=iasesat07

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

The 56th Session of the International Statistical Institute (ISI) will be held in Lisboa, Portugal. IASE is usually very active at ISI meetings (see *SERJ* 4(1) for the report of IASE activities at ISI-55 in Sydney) and will sponsor a list of Invited Paper Meetings (IPMs).

4.1. IPMs SPONSORED BY THE IASE

IPM37 *Research on reasoning about distribution,*
Joan Garfield (jbg@umn.edu)
IPM38 *How modern technologies have changed the curriculum in introductory courses,*
Lucette Carter (lucette.carter@gmail.com)
IPM39 *Preparing teachers of statistics,*
Allan Rossman (arossman@calpoly.edu)
IPM40 *Research on the use of simulation in teaching statistics and probability,*
Rolf Biehler (biehler@mathematik.uni-kassel.de)
IPM41 *Optimizing internet-based resources for teaching statistics* (cosponsored by IASC),
Ginger Holmes Rowell (rowell@mtsu.edu)
IPM42 *Observational studies, confounding and multivariate thinking,*
Milo Schield (milo@pro-ns.net)
IPM43 *Teaching of official statistics* (cosponsored by IAOS) ,
Sharleen Forbes (Sharleen.Forbes@stats.govt.nz)
IPM44 *Teaching of survey statistics* (cosponsored by IASS),
Steve Heeringa (sheering@isr.umich.edu)
IPM45 *Studying variability through sports phenomena* (cosponsored by Sports Statistics),
TBD
IPM46 *Use of symbolic computing systems in teaching statistics* (cosponsored by IASC),
Zaven Karian (Karian@Denison.edu)

4.2. IASE ORGANIZING COMMITTEE:

Allan J. Rossman (USA) arossman@calpoly.edu
Gilberte Schuyten (Belgium) gilberte.schuyten@UGent.be
Chris Wild (New Zealand) c.wild@auckland.ac.nz
For more information visit the ISI 56 website at http://www.isi2007.com.pt/ or contact members of OC.

OTHER FORTHCOMING CONFERENCES

1. JOINT STATISTICAL MEETINGS 2006
Seattle WA, USA, August 6 – 10, 2006

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. Seattle is the host city for JSM 2006 and offers a wide range of possibilities for sharing time with friends and colleagues. For information, contact jsm@amstat.org.

Website: http://www.amstat.org/meetings/jsm/2006/

2. THE 11TH ASIAN CONFERENCE IN MATHEMATICS
ATCM 2006
Hong Kong SAR, China, December 12-16, 2006

2.1. CONFERENCE THEME

The aim of this conference with a theme *Advancing and Fostering Mathematical Sciences and Education through Technology* is to provide a forum for educators, researchers, teachers and experts in exchanging information regarding enhancing technology to enrich mathematics learning, teaching and research at all levels. English is the official language of the conference.

2.2. TOPICS OF INTERESTS

The conference will cover a broad range of topics on the application and use of technology in Mathematics research and teaching. Though Statistics can be recognized in many proposed themes, a special theme *Statistics using Dynamic Statistics Software* might be of special interest for the readers.

Website: http://www.atcminc.com/mConferences/ATCM06/index.shtml

3. THE 6TH ANNUAL HAWAII INTERNATIONAL CONFERENCE ON STATISTICS, MATHEMATICS AND RELATED FIELDS
Honolulu, Hawaii, January 17 – 19, 2007

The 6th Annual Hawaii International Conference on Statistics, Mathematics and Related Fields will be held at the Renaissance Ilikai Waikiki Hotel in Honolulu, Hawaii. The 2007 Hawaii International Conference on Statistics, Mathematics and Related Fields will be the gathering place for academicians and professionals from statistics and mathematics related fields from all over the world.

The main goal of the 2007 Hawaii International Conference on Statistics, Mathematics and Related Fields is to provide an opportunity for academicians and professionals from various statistics and/or mathematics related fields from all over the world to come together and learn from each other. An additional goal of the conference is

to provide a place for academicians and professionals with cross-disciplinary interests related to statistics and mathematics to meet and interact with members inside and outside their own particular disciplines.

Website: http://www.hicstatistics.org/