Statistics Education Research Journal

Volume 4 Number 1 May 2005

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

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

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. Manuscripts submitted before 1 November 2005 should be sent to co-editor Flavia Jolliffe <F.Jolliffe@kent.ac.uk>. Manuscripts submitted after 1 November 2005 should be sent to co-editor Iddo Gal <iddo@research.haifa.ac.il>. 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, Voorgurg, The Netherlands
Technical Production: University of New England, Armidale, NSW, Australia

ISSN: 1570-1824

TABLE OF CONTENTS

EDITORIAL

SERJ is in its fourth year of operation and it is clear that it is now well established as the research journal of the International Association for Statistical Education (IASE). The flow of new manuscripts, as well as their breadth, are increasing, and represent the growing interest in research and in new knowledge that can inform practice in statistics education. That said, many areas of importance for statistics education are underrepresented in current research, such as: learning about associations and correlations, learning advanced topics such as regression or inference testing, the link between knowledge of probability and learning of statistical inference, students' ability to apply and transfer knowledge to out-of-school situations which require activation of statistical understanding, or factors that affect and programs that can improve adults' understanding of real-world statistical messages and arguments.

The selected examples given above are far from being exhaustive; they are provided merely to illustrate the range of research areas that have a potential to contribute to improvement of statistics learning, teaching, and application by people in different educational, cultural and functional contexts. We encourage researchers and educators from diverse disciplines to collaborate, and to consider expanding and extending research plans, in order to address the research and practice needs of the international statistics education community.

This issue contains six papers, four of which appear in a special section on research on variation, which extends the special issue we published in November 2004 on that topic. We thank Joan Garfield and Dani Ben Zvi, who were the Guest Editors both for the former special issue as well as for the special section in the current issue, for their work and initiative regarding this important area.

The four papers on variation include two refereed research papers (Makar & Confrey; delMas & Liu) and two invited discussion papers (Garfield & Ben-Zvi; Pfannkuch). Katie Makar and Jere Confrey examine how prospective secondary mathematics and science teachers understand and articulate notions of variation as they compare two distributions. Bob delMas and Yan Liu examine students' understanding of the standard deviation, and the impact of using a customized computer applet, on their reasoning about the link between spread and center. Maxine Pfannkuch discusses broad implications of the empirical papers on reasoning about variation published in the November 2004 special issue, with an emphasis on the role of tools in students' learning and in future research, and the link between learning about variation and broader aspects of the statistical enquiry cycle. Joan Garfield and Dani Ben-Zvi further extend their reflection based on the empirical papers on reasoning about variation, by pointing to a model that can inform instruction, assessment, and future research. This issue also includes two regular research papers. Linda Brant Collins and Kathleen Mittag deal with the use of calculators in teaching statistics and their paper fills an important gap in the literature. Elena Papanastasiou describes a scale developed to measure "attitudes towards research" of college students, thus adding to the literature that so far has focused more on attitudes and beliefs regarding statistics.

We now turn to a brief report of changes and future plans at SERJ.

First, we plan a special issue on research on "learning and teaching of reasoning about distributions" for November 2006. A preliminary announcement was circulated a few months ago, and a more detailed Call for Papers appears later in this issue. The deadline for submissions is 1 November 2005. Interested authors are asked to submit a letter of intent and to follow the guidelines in the Call for Papers.

Second, there have been some recent changes to our editorial board. Three associate editors have departed, Carmen Batanero, Annie Morin, and Chris Wild. We thank all three for the many

Statistics Education Research Journal, 4(1), 2-3, http://www.stat.auckland.ac.nz/serj

contributions they have made to the development of SERJ while serving on the board. Carmen was a founding Editor and was instrumental in the transition from the former *Statistical Education Research Newsletter*. Chris was also involved with SERJ from the start and Annie joined soon after. Chris, while president of IASE, developed the Association's web pages and SERJ's web page is naturally part of that site. We welcome two new associate editors, Gilberte Schuyten from Belgium, and Ernesto Sanchez from Mexico. Biographical notes for both are given on the next page.

Third, Flavia Jolliffe's four-year term as co-editor ends on 31 December 2005. The search for a new co-editor is progressing, following a procedure recently formalised by the IASE Executive. A 3-person search committee is being formed, consisting of a member of the IASE Executive (chair) appointed by the IASE president, the continuing co-editor, and a member-at-large of IASE who is neither on the IASE Executive nor on the SERJ editorial board. A Call for Nominations is published later in this issue as well as on the IASE website under 'Publications'.

Fourth, we continue updating the guidelines for authors and other SERJ documentation. We expect the revised guidelines to be available in July 2005 and encourage prospective authors to examine these materials and follow them in future submissions. We take this opportunity to express our gratitude to Chris Reading, SERJ's Assistant Editor, for the many hours she puts in, and the care she takes, in producing SERJ to a high professional standard.

Finally, now that plans are well underway for ICOTS7 in 2006 and for many other conferences where research papers in statistics education are presented, we would like to remind prospective authors to be attentive to "prior publication" or "duplicate publication" policies which different journals apply. Like many journals, SERJ's policy is that papers already published, i.e., available for wide public consumption via the Internet or other electronic or printed means, cannot be accepted for consideration by SERJ. In the case of submissions based on papers previously published in conference proceedings, whether in print or electronically, we expect that submitted papers will be *substantially different or expanded*. This usually does not present a problem as many conferences typically pose a limit on word/page count, so what is published is limited in scope from the outset. The upshot is that authors have to strategize in advance what selected portions they submit for publication in conference proceedings and what additional materials, results, analyses, and discussions will be added and be exclusive to manuscripts submitted for journal consideration and review. As will be explained in our revised guidelines, authors will be expected to declare upon submission if a paper or a portion of it was previously published in any form. Authors are encouraged to consult the editors in advance if doubts exist as to what constitutes prior publication, in order to maximize the match of author intentions and journal expectations, and make sure authors find a suitable outlet for their research work.

In closing, we reiterate our conviction that the Journal is supposed to serve a diverse and expanding community of practitioners and researchers interested in statistics education and learning in diverse fields and contexts. We thus encourage all readers of SERJ to send us comments and suggestions regarding the journal, its scope, papers it publishes, and ideas for future plans.

IDDO GAL AND FLAVIA JOLLIFFE

NEW ASSOCIATE EDITORS

SERJ welcomes the following new Associate Editors, who have joined the Editorial Board for a 3-year appointment 2005-2007.

Ernesto Sanchez has a Ph. D. in Mathematics Education with a further background in mathematics. He is a Researcher at the Center for Research and Advanced Studies of 'Instituto Politécnico Nacional' in Mexico. His research interests have focused on topics of teaching and learning of probability such as stochastic independence, conditional probability, and relationships between probabilistic thinking and technology. He has published numerous research articles in statistics education in the Spanish language. Recently he was a coauthor with Carmen Batanero of a chapter 'What is the nature of high school student's conceptions and misconceptions about probability?' in the book by Graham Jones (2005) on research and teaching of probability in schools.

Gilberte Schuyten is Professor and Head of the Department of Data Analysis at the Faculty of Psychology and Educational Sciences at Ghent University, Belgium. She teaches data analysis and empirical research methods at the Faculty. Her Ph.D. is in Mathematics from Ghent University (1971). As a young researcher she started 'new math' experiments in the late sixties, and later her research interests centered on technology and the school curriculum. She introduced Logo in Belgium, organized training courses for Flemish teachers, and directed projects of the Flemish government aimed at using ICT in statistics courses at universities. At the European level she has organized international meetings and conferences about information technologies at school. Her primary interest is the influence of cognitive and affective characteristics and the use of electronic technologies on statistics learning of students in the social sciences.

Statistics Education Research Journal, 4(1), 4, http://www.stat.auckland.ac.nz/serj

CALL FOR NOMINATIONS FOR NEW CO-EDITOR: STATISTICS EDUCATION RESEARCH JOURNAL

The International Association for Statistical Education (IASE) is starting a search for the next co-editor of *Statistics Education Research Journal* (SERJ), its peer-reviewed electronic journal. The new editor will serve a four-year term starting Jan 1, 2006, replacing Flavia Jolliffe (U. of Kent, UK), who will end her four-year tenure in Dec 2005. The new editor will join Iddo Gal (U. of Haifa, Israel), the continuing co-editor until Dec 2007.

SERJ was established in 2002 by IASE 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 and informal contexts. The breadth and scope of manuscripts submitted to SERJ are increasing and represent the growing interest in research and in new knowledge that can inform practice in statistics education. The SERJ organization includes two co-editors who serve for 4 years (one is replaced every two years), an Assistant Editor in charge of copy-editing and production, and an Editorial Board presently comprised of 12 Associate Editors from 10 countries. SERJ issues and materials are published on the IASE website, presently hosted by the University of Auckland (NZ). The journal maintains autonomy regarding content and process, although some activities are coordinated with IASE and its parent organization and co-publisher, the International Statistical Institute (ISI). All journal activities are conducted electronically. Board members meet during key international conferences such as ICOTS or ISI. SERJ is a virtual organization and it operates on the basis of voluntary work by all board members and editors.

The co-editors are responsible for overall management of all journal operations, determine the composition of the Editorial Board and the reviewer pool, initiate and conduct communication with prospective authors, reviewers, associate editors, and external stakeholders, and manage peer-review and editorial processes. The co-editors are expected to establish editorial policies, set scholarly and quality expectations, and uphold acceptance criteria regarding manuscripts. The co-editors should have a forward-looking vision and initiate new features and structures, if needed in consultation with Board members and others, so as to enable SERJ to respond to the evolving knowledge needs in the dynamic area of statistics education. Overall, the co-editors should lead the journal to make an important contribution to research and practice in statistics education.

The qualified individual will have a research and practice background of relevance to statistics education, possess the skill to work with prospective contributors in a supportive yet critical spirit, and be able to maintain and strengthen international professional networks of authors and reviewers and enhance the Journal's reputation and impact.

The search process and how to apply

IASE encourages both nominations of suitable candidates and self-nominations from interested individuals. All nominations and self-nominations will be considered by the Search Committee, which can also propose additional nominees. Candidates or self-nominees will be asked to prepare a brief statement describing their vision for continuing the growth and development of the Journal, along with a statement of their qualifications for the position, and an academic vita or professional resume. Candidates might also be asked to respond to additional questions by the search committee.

For more information about the search process, or to submit a nomination, please contact the Chair of the Search Committee, Chris Wild, (U. of Auckland. NZ): <c.wild@auckland.ac.nz>. Questions about the practicalities of the editorship can be sent to either the continuing co-editor, Iddo Gal <iddo@research.haifa.ac.il> or to the departing co-editor, Flavia Jolliffe: <F.Jolliffe@kent.ac.uk>. Nominations should be submitted as soon as possible, preferably not later than July 15, 2005. The editorial change is expected to occur January 1, 2006.

CALL FOR PAPERS: REASONING ABOUT DISTRIBUTIONS

The *Statistics Education Research Journal* (SERJ), a journal of the International Association for Statistical Education (IASE), is planning a special issue for November 2006, focused on research on reasoning about distributions. Submission Deadline: November 1, 2005.

The aim of the special issue is to advance the current state of research-based knowledge about the learning and teaching of reasoning about distributions, and to contribute to future research and to research-based practice in this area. Little research, whether qualitative or quantitative, has been published to date in this area, despite “distribution” being a foundational topic in statistics and one of the underpinnings of statistical literacy. Many research challenges exist, such as regarding knowledge of students and educators in diverse contexts of learning distribution-related ideas (e.g., K-12, tertiary, workplace), effective curricular materials and tools, or methods for documenting knowledge or measuring performance on tasks that require understanding of distributions.

Examples for relevant topics for research-oriented papers include (but are not limited to):

- How students or adults understand distributions, or make use of information about distributions, whether as a stand alone topic or in relation to reasoning about other statistical topics or tasks (e.g., involving variation, statistical inference, probability),
- How technological tools are utilized by learners to generate representations or improve thinking about distributions,
- What developmental trajectories exist, e.g., in acquisition of informal and formal knowledge about distributions, in learning to represent distributions, in proficiency in interpreting information or displays about distributions,
- How students interpret information regarding distributions when generated by technology or other means, and how these interpretations can be improved,
- What misconceptions and difficulties can be seen when students or adults think about or work with information about distributions, and what may be their origins,
- How does learners’ knowledge of distributions, or difficulties they encounter in this regard, contribute to or impede their behavior and thinking when coping with tasks involving other topics in statistics and probability,
- Knowledge and perspectives of educators involved in teaching about distributions,
- The relative efficacy of teaching approaches or curricular materials that can promote the understanding of distributions or their use in various tasks,
- Innovative assessment approaches and research methodologies in this area.

Manuscripts will be limited to a maximum of 8500 words of body text (not counting abstract, tables and graphs, references, appendices). Shorter, concise papers are preferred. Manuscripts will be reviewed following SERJ’s regular double-blind refereeing process. Guest Editors of this special issue will be Maxine Pfannkuch (University of Auckland, New Zealand) and Chris Reading (University of New England, Australia).

Interested authors are asked to send a letter of intent with details of the planned paper, or any queries, to Iddo Gal, SERJ co-editor, at: <iddo@research.haifa.ac.il>. Manuscripts must be submitted by November 1, 2005 to the same address, using the SERJ Author Guidelines and Template found on: <www.stat.auckland.ac.nz/serj>. (Please be advised that the Author Guidelines and Template will be updated in July 2005.)

Statistics Education Research Journal, 4(1), 6, http://www.stat.auckland.ac.nz/serj

EFFECT OF CALCULATOR TECHNOLOGY ON STUDENT ACHIEVEMENT IN AN INTRODUCTORY STATISTICS COURSE

LINDA BRANT COLLINS
University of Chicago
collins@galton.uchicago.edu

KATHLEEN CAGE MITTAG
University of Texas at San Antonio
kmittag@utsa.edu

SUMMARY

We report on a study of the relationship between calculator technology and student learning in two introductory statistics class sections taught by the same instructor at the University of Texas at San Antonio. At the introduction of hypothesis testing and confidence intervals, one class section (A) was given graphing calculators capable of inferential statistics to use for a few weeks. At the same time, the other class section (B) was given non-inferential graphing calculators. Data were collected on all test grades and daily quiz grades for both class sections. The students were allowed to use the inferential calculators on only the examination covering hypothesis tests and confidence intervals and on the final examination. Both sections received the same tests. We found that although use of the calculator with inferential capabilities is associated with improved scores on the inferential examination, the improvement is not significant once we adjust for performance on previous tests. Still, we note that on final examination questions specifically utilizing the calculator inference functions, the two classes perform similarly. In fact, both classes had trouble with "calculations" while at the same time answering "concept" questions fairly well. The inferential calculator did not appear to give students any clear advantage or disadvantage in their performance on examinations.

Keywords: Statistics education research; Introductory statistics; Graphing calculator; Inferential calculator; Student achievement

1. INTRODUCTION

Since calculators with inferential statistics capabilities came on the market in January, 1996, it has become evident that statistics educators need to analyze the effectiveness of the new hand-held technology. It is interesting to note that many statistics instructors at our university were not aware of these calculators and this may also be true at many other universities.

We studied the effect of calculator technology on student achievement in two introductory statistics class sections taught by one of the authors in autumn, 1998, at a large public urban university in the United States. At the introduction of the topics of hypothesis testing and confidence intervals, one class section (Class A) was given inferential calculators to use for a few weeks. At the same time, the other class section (Class B) was given older calculators without inferential capabilities. Other than this difference in calculators, the two groups were treated as similarly as possible.

Statistics Education Research Journal, 4(1), 7-15, http://www.stat.auckland.ac.nz/serj

In this paper, we first review the current literature on technology in the statistics (and mathematics) classroom and then proceed to an analysis of the data collected from our own study.

2. BACKGROUND: COMPUTERS

Many efforts are being made to enhance the learning experiences for students in introductory statistics courses (Cobb, 1993; Garfield, 1995; Gnanadesikan, Scheaffer, Watkins, & Witmer, 1997). Technology is influencing the teaching and learning of statistics. Gilchrist (1986) suggested that computers should be utilized to teach concepts and methods and that number crunching should be de-emphasized. Singer and Willet (1990) asserted that since the advent of computers, artificial data sets are no longer needed. David Moore (1992) suggested that calculation and graphics be automated as much as possible. Hogg (1992), Neter (1989) and Pirie (1989) wrote that the use of computers in undergraduate statistics classrooms is very important. A computer-based instructional strategy should be used either for managing large data sets or for generating simulations to illustrate probability concepts (Mittag & Eltinge, 1998).

Moore wrote in Moore, Cobb, Garfield, and Meeker (1995) that

> The nature of both statistical research and statistical practice has changed dramatically under the impact of technology. Our teaching has certainly changed as well --- but what strikes me is how little it has changed. The computing revolution has changed neither the nature of teaching nor our productivity as teachers. (p. 250)

Moore goes on to suggest that some reasons for educators being slow to change may be: “our costs have risen much faster than incomes or inflation; we see no need to change; we have an outdated organizational structure; and there is little internal incentive to change” (p. 251).

3. BACKGROUND: CALCULATORS

The advent of calculator technology has influenced the teaching of mathematics in a profound way (Dunham & Dick, 1994; Demana & Waits, 1990; Fey & Good, 1985). Several research studies have documented the benefits of calculator use in the mathematics classroom (Campbell & Stewart, 1993; Carlson, 1995; Dunham, 1993; Dunham, 1996; Graham & Thomas, 2000; Harvey, Waits, & Demana, 1995; Hembree & Dessart, 1986; Hennessy, Fung, & Scanlon, 2001; Quesada & Maxwell, 1994). Research has not focused on the use of calculator technology in the statistics classroom. There have been a few papers published which discuss graphing calculators in statistics. Rinaman (1998) discussed changes that had been made in basic statistics courses at his university. The TI-83 graphing calculator was recommended for the students to purchase but it was determined that all students should be required to purchase this or no calculator should be allowed at all. Garfield (1995, p.31) wrote that “calculators and computers should be used to help students visualize and explore data, not just to follow algorithms to predetermined ends.” A sample lesson using the TI-80 calculator based on modeling and simulation was discussed by Graham (1996a). Binomial graph and Poisson graph programs for the TI-82 were presented and demonstrated by Francis (1997). Statistical features of the TI-83 and TI-92 were discussed by Graham (1996b). Since there has been little, if any, research on the effect of graphing calculators on conceptual understanding in introductory statistics, the authors decided to conduct the study described in this paper.

4. EXPLANATION OF CALCULATORS USED IN THE STUDY

Statistics students have used the scientific calculator for the past two decades. With the introduction of the graphing calculator about ten years ago, basic descriptive statistics and graphing were automated. One or two variable data sets ($n < 100$) could be entered in the calculator then:

descriptive statistics, such as the mean and standard deviation, calculated; graphs such as histogram and scatterplot, displayed; and some regression equations, such as linear, exponential, ln, log, power and inverse, could be calculated. In our study, a graphing calculator with the above capabilities was furnished to Class B.

Class A received a graphing calculator with inferential statistics capabilities. At the time of writing, Casio, Sharp and Texas Instruments offer these calculators for less than \$100 in the USA. These calculators have many sophisticated statistical capabilities that include inferential statistics such as hypothesis testing, confidence interval calculations, and one-way analysis of variance.

5. WHY USE THE GRAPHING CALCULATOR?

There are several reasons to use the advanced graphing calculator in an introductory statistics course. The major reasons are access and economics (Kemp, Kissane, & Bradley, 1998). For an investment of less than \$100, a student has access to technology at home and in the classroom at all times. A student can do homework and take examinations using the calculator, which is difficult to do with a computer. Many students already own a calculator, especially if they are recent high school graduates. The Advanced Placement examinations in Calculus and Statistics require the use of a graphing calculator. The high school mathematics curriculum has included the incorporation of calculator technology for several years. Of course, the calculator does not have all the capabilities of computer software packages. Data set size and analyses are limited and printing results is not as easy. However, the calculator is a useful tool and teaching aid for introductory statistics.

6. BASIC METHODOLOGY OF THE STUDY

Class A consisted of 22 individuals who completed one section of an introductory statistics course and were provided with a calculator capable of inferential statistics. Class B was 47 individuals who completed another section of the same introductory statistics course with the same instructor. These students were also provided with a calculator, but without the facility for direct inferential statistics. Students enrolled in either Class A or Class B on their own, with no knowledge of the existence of the study. Students in both sections of the course were told that they were being asked to use the calculator in an effort to assess its effectiveness and all students used the calculators on the analyzed examinations. The instructor used the non-inferential calculator overhead while lecturing to both sections. Students were expected to show all the traditional calculations on the inferential examination. The teacher demonstrated the inferential capabilities of the calculator in Class A and did not discuss the inferential capabilities in Class B. Both classes had instruction on other similar statistical capabilities of the two calculator models. The students from each class did not meet or work together to discuss instruction. During examinations, the teacher made sure that the correct calculators were being used. She did not allow the inferential calculator to be used in Class B. The final examination was 2 hours and 45 minutes long and all the problems were compulsory. Both class sections had the same examination though there were four different forms. (See the appendix for an example question with answers.) Every effort was made to keep the experience and evaluation of the two groups as similar as possible.

As an example of the difference in capabilities of the two calculators, consider the exercise of constructing a confidence interval for a proportion, p. Students using the inferential calculator would simply input the confidence level, number of observed successes, and total sample size. Then, the calculator reports an estimate and confidence interval for p. Students using non-inferential calculators would need to first calculate the sample proportion, estimate the standard error, find the appropriate z-value to create the margin of error, and then find the two endpoints of the confidence interval.

For a hypothesis test on a proportion, p, students using the inferential calculator would input the null hypothesis, number of observed successes, total sample size, and type of tail for the test. Then, the calculator reports the test statistic and the p-value for the test. Students using the non-inferential calculators would need to first calculate the sample proportion, estimate the standard error, create the test statistic, and then look up the critical value and p-value for the test. One sample inference problem from the final examination with answers is given in the appendix.

7. DESCRIPTION OF THE DATA

The following variables were collected from the 69 students who completed one of two sections of an introductory statistics class during one spring semester at the university:

- Gender, ethnicity, major;
- Pretest scores (test scores on the first three examinations);
- Inferential test score (test score on the fourth examination);
- Daily grade (a score based on 20 daily quizzes and homework converted to a 100-point scale);
- Final examination score (and answers to individual final examination questions).

All individual examination scores and the daily grade were recorded on a 100-point scale.

Both classes occurred on a Monday/Wednesday afternoon schedule (Class A at 3:30pm and Class B at 2:00pm) with the same instructor and the same examinations. Unfortunately, many variables (both observed and unobserved) confound the study and we cannot separate the “classroom effect” from the “calculator effect.” In fact, we observed several differences between Class A and Class B besides the assignment of different calculators for the study. For example, Class A was smaller (22 students) than Class B (47 students) and Class A suffered a greater withdrawal rate of 8 students compared to just 1 withdrawal from Class B. Also, those students remaining in Class A for the entire semester scored significantly higher on their examinations both before and after receiving the calculators for the study.

Nonetheless, there are interesting facets in the data. For example, although students in Class A tended to score higher on both examination 3 (prior to receiving the inferential calculator) and examination 4 (after getting the calculator), they did not show any greater “improvement” in performance after receiving the calculator. We demonstrate these phenomena in the analysis that follows.

Note that the 63 students included in the data analyses (21 from Class A and 42 from Class B) represent only those students who completed the final examination. In addition, sample sizes for the other examinations vary slightly since students were allowed to drop one test score (not the final examination) and indeed three students did not complete examination 4 (one student from Class A and two from Class B). These students provided no information about the effect of an inferential calculator on their examination 4 performance.

8. DATA ANALYSIS

Table 1 gives some summary information on the two groups of students in this study. Recall that students in Class A received an inferential calculator after examination 3 and students in Class B received a non-inferential statistical calculator as described in Section 6. The t-test p-values are for the two-sided pooled-variance t-test of the difference in mean scores. Bartlett’s test of the equality of variances (not shown) indicates that the variance of scores are not significantly different for the two classes for the various examinations (except examination 4). The t-test for examination 4 uses Satterthwaite’s approximate degrees of freedom. Dotplots (not shown) of the test scores show no significant departure from an assumption of normality.

Students using the inferential calculator scored an average of 10.1 points higher (on a 100-point scale) on the inferential statistics examination (examination 4) than those students using the non-inferential calculator. The two-sided p-value for an unequal-variance t-test of the difference in mean examination 4 scores is 0.03. However, since we also have information on each student’s general test-taking ability from the previous examination (examination 3), we can examine the effect of the calculator on examination scores while controlling for a student's examination-taking ability. The plot of the data in Figure 1 illustrates the relationships. In general, students scoring higher on examination 3 also score higher on examination 4.

Table 1. Data Summary

Description	Class A	Class B	
Number of Students	21	42	
Female Students	12 (57%)	20 (48%)	
3rd/4th-year Students	12 (57%)	18 (43%)	
Hispanic Surname	13 (62%)	22 (52%)	
	Avg (sd)	Avg (sd)	t-test p-value
Examination 3	80.0 (16.3)	71.1 (14.3)	0.03
Examination 4	82.3 (13.7)	72.2 (21.3)	0.03
Final Examination	77.2 (10.5)	74.8 (10.3)	0.38
Daily Grade	85.0 (17.7)	85.2 (18.4)	0.97

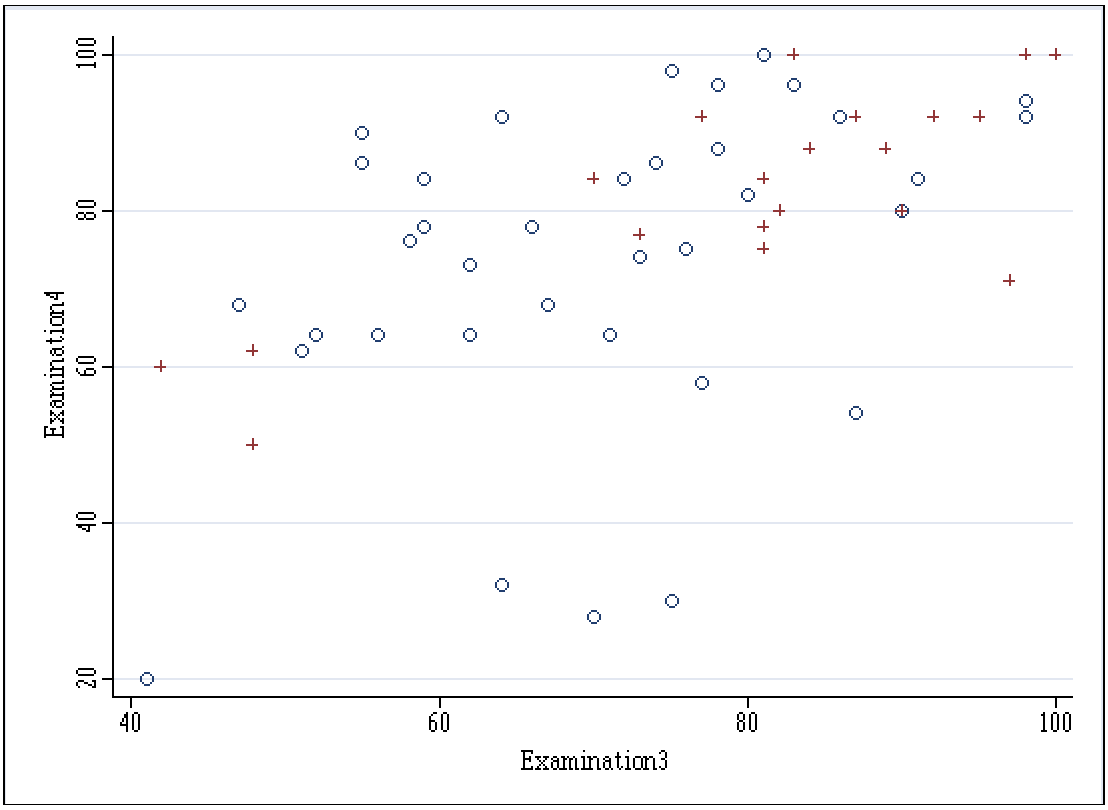


Figure 1: Examination 4 scores by Examination 3 scores. Class A (students using the inferential calculator) marked as "+"

The following linear regression model was fitted to the data:

Examination4 = Intercept + E*Examination3 + C*Calculator.

Here,

Examination4 = score on Examination 4 (inferential topics),

Calculator = 1 for Class A (inferential calculator) and 0 for Class B,

Examination3 = score on Examination 3 (test score prior to obtaining calculators for the study).

The regression results are recorded in Table 2.

Table 2. Least squares regression results to estimate calculator effect while adjusting for prior examination scores

Dependent Variable: examination4

Coefficient	Parameter Estimate	Standard Error	T for Ho: Parameter=0	P-Value: Prob > |T|
Intercept	33.640	12.210	2.76	0.008
E (Examination3)	0.608	0.146	4.17	0.000
C (Calculator)	2.690	4.716	0.57	0.571

R-Square = 28.34%

The examination 3 scores are, as expected, a strongly significant predictor of the average examination 4 score (p-value < 0.0001). However, once the model is adjusted for this measure of examination performance (examination 3 score), the type of calculator used is not at all statistically (or practically) significantly related to performance on the inferential statistics examination (examination 4). A similar analysis with the final examination score as the response variable indicates no significant difference between the two groups of students either conditionally (adjusting for previous scores on examination 3) or unconditionally (as already seen in Table 1 where the p-value for a t-test of the difference in average final examination scores was 0.38).

We should point out that, generally, there is a lot of variability in these data and even the examination 3 scores only account for about 28% of the variability in examination 4 scores when using the linear regression model fit in Table 2. The relationship between examination 3 and examination 4 scores appears fairly linear for both groups, but there is a lot of scatter in the data. A plot of the residuals does not indicate the presence of a non-linear pattern.

To look further for any effect of the inferential calculator, we took a deeper look at students' performance on specific examination questions. See, for example, the sample final examination question in the Appendix. In this question, Parts I, II, and V could be considered conceptual and III, IV, and VI calculations (with some overlap). We found that both classes had trouble with the "calculations" while at the same time answering "concept" questions fairly well.

9. DISCUSSION AND CONCLUSIONS

In general, we observed no difference between the two groups of students in their performance on examinations on inferential topics in an introductory statistics course. In particular, use of an inferential calculator that performs many of the intermediate steps for calculating confidence intervals and p-values does not appear to be related to student performance. The study size was small and the design did not allow for the separation of the "calculator effect" from the "classroom effect" (a confounding factor). However, it is interesting to note that although Class A generally performed better on examinations prior to receiving inferential calculators for the study, these same students did not perform significantly or practically better on inference topics after using the calculator. The inferential calculator did not appear to give students any clear advantage (or disadvantage) in their performance on examinations. This study suggests that use of the inferential calculator needs to be explored further as a benefit for student-learning in an introductory statistics courses. We encourage others who may teach within an infrastructure that could allow a randomized experiment to complete a study and share their results.

Of course, if the inferential calculator is required in an introductory statistics course, the instructor would be able to spend much less time on computation and more time on gathering and analyzing real-world data. Indeed, as classroom computer use has changed how statistics has been taught during the last 20 years (as computers have changed the practice of statistics), inferential calculators can also change how introductory statistics is taught.

ACKNOWLEDGEMENTS

This research was completed at the University of Texas at San Antonio.

REFERENCES

Campbell, P., & Stewart, E.L. (1993). Calculators and computers. In R. Jensen (Ed.), *Early Childhood Mathematics, NCTM Research Interpretation Project* (pp. 251-268). New York: Macmillan Publishing Company.

Carlson, M. P. (1995). A successful transition to a calculator integrated college algebra curriculum: Clues, surveys, and trends. In P. Bogack (Managing Ed.), E. D. Fife, & L. Husch (Eds.), *Proceedings of the Seventh Annual International Conference on Technology in Collegiate Mathematics* (pp. 73-77). Reading, MA: Addison-Wesley Publishing Group.

Cobb, G. W. (1993). Reconsidering statistics education: A National Science Foundation conference, *Journal of Statistics Education, 1*(1). [Online: www.amstat.org/publications/jse]

Demana, F., & Waits, B. K. (1990). Implementing the Standards: The role of technology in teaching mathematics, *Mathematics Teacher*, *83*, 27-31.

Dunham, P. H. (1993). Does using calculators work? the jury is almost in. *UME Trends*, *5*(2), 8-9.

Dunham, P. H. (1996). Looking for trends: What's new in graphing calculator research? In P. Bogack (Managing Ed.), E. D. Fife, & L. Husch (Eds.), *Proceedings of the Eighth Annual International Conference on Technology in Collegiate Mathematics* (pp. 120-124). Reading, MA: Addison Wesley Publishing Company.

Dunham, P. H., & Dick, T. P. (1994). Research on graphing calculators. *Mathematics Teacher*, *87*, 440-445.

Fey, J. T., & Good, R. A. (1985). Rethinking the sequence and priorities of high school mathematics curricula. In C. R. Hirsch & J. Zweng (Eds.), *The secondary school mathematics curriculum, 1985 yearbook* (pp. 43-2). Reston, VA: National Council of Teachers of Mathematics.

Francis, B. (1997). Graphs of binomial and Poisson distributions on a graphical calculator. *Teaching Statistics*, *19*(1), 24-25.

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

Gilchrist, W. (1986). Teaching statistics to the rest of humanity. In R. Davidson & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 494-497). Victoria, British Columbia, Canada: University of Victoria Conference Services.

Gnanadesikan, M., Scheaffer, R. L., Watkins, A. E., & Witmer, J. R. (1997). An activity-based statistics course. *Journal of Statistics Education, 5*(2).

Graham, A. (1996a). Following form with a graphic calculator. *Teaching Statistics*, *18*(2), 52-55.

Graham, A. (1996b). The TI-83 and TI-92 calculators. *Teaching Statistics*, *18*(3), 94-95.

Graham, A. T., & Thomas, M. O. J. (2000). Building a versatile understanding of algebraic variables with a graphic calculator. *Educational Studies in Mathematics*, *41*(3), 265-282.

Harvey, J. G., Waits, B. K., & Demana, F. D. (1995). The influence of technology on the teaching and learning of algebra. *Journal of Mathematical Behavior*, *14*, 75-109.

Hembree, R., & Dessart, D. J. (1986). Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *Journal for Research in Mathematics Education*, *17*, 83-89.

Hennesey, S., Fung, P., & Scanlon, E. (2001). The role of the graphic calculator in mediating graphing activity. *Education in Science and Technology*, *32*, 267-290.

Hogg, R.V. (1992). Towards lean and lively courses in statistics. In F. Gordon & S. Gordon (Eds.), *Statistics in the twenty-first century [MAA Notes No. 26]* (pp. 3-13). Washington, DC: Mathematical Association of America.

Kemp, M., Kissane, B., & Bradley, J. (1998). Learning undergraduate statistics: The role of the graphics calculator. In B. Parker (Ed.), *Proceedings of the International Conference on the Teaching of Mathematics* (pp. 176-178). Reading, MA: John Wiley and Sons, Inc.

Mittag, K. C., & Eltinge, E. (1998). Topics coverage in statistics courses: A national Delphi study. *Research in the Schools*, *5*(1), 53-60.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. Gordon, & S. Gordon (Eds.), *Statistics in the Twenty-First Century [MAA Notes, No. 26]* (pp. 14-25). Washington, DC: Mathematical Association of America.

Moore, D. S., Cobb, G. W., Garfield, J., & Meeker, W. (1995). Statistics education fin de siecle. *The American Statistician*, *49*(3), 250-260.

Neter, J. (1989). Undergraduate statistics service courses in the years ahead. In J. Bolinger & R. Brookmejer (Eds.), *American Statistical Association, 1989 Proceedings of the Section on Statistical Education* (pp. 29-31). Alexandria, VA: American Statistical Association.

Pirie, W. R. (1989). Undergraduate statistics programs: Challenges for the 1990's. In J. Bolinger & R. Brookmejer (Eds.), *American Statistical Association 1989 Proceedings of the Section on Statistical Education* (pp. 32-33). Alexandria, VA: American Statistical Association.

Quesada, A. R., & Maxwell, M. E. (1994). The effects of using graphing calculators to enhance college students' performance in precalculus. *Educational Studies in Mathematics*, *27*, 205-215.

Rinaman, W. C. (1998). Revising a basic statistics course, *Journal of Statistics Education*, *6*(2). [Online: www.amstat.org/publications/jse]

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

LINDA BRANT COLLINS
Department of Statistics
University of Chicago
5734 S. University Ave
Chicago, IL 60637
United States of America

APPENDIX: FINAL EXAMINATION INFERENCE QUESTION WITH ANSWERS

A statistics professor surveys 60 students and finds that 12 are left-handed. Use a 0.05 level of significance to test the claim that these students come from a population in which the percentage of left-handed people is greater than 10%.

Part I: What is the correct null and alternative?

a) H0: $p \leq 0.10$ H1: $p > 0.10$

b) H0: $p > 0.10$ H1: $p \leq 0.10$

c) H0: $p = 0.10$ H1: $p \neq 0.10$

d) H0: $p \neq 0.10$ H1: $p = 0.10$

e) none of these

Answer: a) H0: $p \leq 0.10$ H1: $p > 0.10$

Part II: Which of the following is true?

a) This is a two-tailed test.

b) This is a left-tailed test.

c) This is a right-tailed test.

Answer: c) This is a right-tailed test.

Part III: What is the test statistic?

a) -1.94

b) -2.58

c) 2.58

d) 1.94

e) none of these

Answer: c) 2.58

Part IV: What is the critical value?

a) 1.645

b) 2.575

c) 1.96

d) none of these

Answer: a) 1.645

Part V: What is the conclusion?

a) There is not sufficient evidence to reject the claim that the proportion is more than 0.10.

b) There is not sufficient evidence to support the claim that the proportion is more than 0.10.

c) There is sufficient evidence to reject the claim that the proportion is more than 0.10.

d) There is sufficient evidence to support the claim that the proportion is more than 0.10.

e) none of these

Answer: d) There is sufficient evidence to support the claim that the proportion is more than 0.10.

Part VI: What is the p-value for this problem?

a) 0.0049

b) 0.4951

c) 0.4738

d) 0.0262

e) none of these

Answer: a) 0.0049

FACTOR STRUCTURE OF THE "ATTITUDES TOWARD RESEARCH" SCALE

ELENA C. PAPANASTASIOU
Intercollege, Cyprus
elenapap@ucy.ac.cy

SUMMARY

Students at the undergraduate level usually tend to view research methods courses negatively. However, an understanding of these attitudes is necessary to help instructors facilitate the learning of research for their students, by enabling them to create more positive attitudes toward such courses. The aim of this study is to describe the development of an "attitudes toward research" scale and verify the dimensions of attitudes toward research among undergraduate students enrolled in introductory research courses. The basic hypothesis of this research study is that the concept of attitudes is multidimensional in nature. The sample of the study consisted of 226 students who had completed a research methods course. Based on a factor analysis, five factors of student attitudes toward research were identified. These were the factors of usefulness of research, anxiety, affect indicating positive feelings about research, life relevancy of research to the students' daily lives, and difficulty of research.

Keywords: Statistics education research; Research methods; Quantitative research attitudes; Scale development; Factor structure; Attitudes toward research; Psychometrics

1. THEORETICAL FRAMEWORK

Students at the undergraduate university level, typically tend to view research-related courses with negative attitudes and feelings. These negative attitudes have been documented in numerous studies for a number of years in relation to courses in research, statistics and mathematics (Adams & Holcomb, 1986; Elmore & Vasu, 1980; Wise, 1985). One of the main problems of these negative attitudes is that they have been found to serve as obstacles to learning (Wise, 1985; Waters, Martelli, Zakrajsek, & Popovich, 1988). In turn, these negative attitudes have been found to be associated with poor performance in such courses (Elmore, & Lewis, 1991; Woelke, 1991; Zeidner, 1991). Causal models, however, suggest that attitudes are actually mediators between past performance and future achievement (Meece, Wigfield, & Eccles, 1990).

Prior research studies have found that negative attitudes toward a course (e.g., mathematics) have been found to explain a significant portion of the variance in student learning (Ma, 1995). In turn, these attitudes influence the amount of effort one is willing to expend on learning a subject, which also influences the selection of more advanced courses in similar areas (e.g., research and statistics courses) beyond those of minimum requirements. Therefore, assessing students' attitudes toward a research methods course is important in order to enable instructors to develop instructional techniques leading to more positive attitudes toward the subject (Waters et al., 1988).

In a 1980 study, Roberts and Bilderback (1980) found that most students who take statistics are quite anxious. Once this preponderance of negative attitudes was revealed, many more survey instruments designed to measure university students' attitudes toward statistics were developed (Dauphinee, Schau, & Stevens, 1997; Zeidner, 1991). One such instrument is the Survey of 'Attitudes Towards Statistics' (Schau et al., 1995), which is comprised of four dimensions, those of affect, cognitive competence, value, and attitudes about the difficulty of statistics. Another instrument

Statistics Education Research Journal, 4(1), 16-26, http://www.stat.auckland.ac.nz/serj

created for the same purpose was that of Attitudes Toward Statistics (Wise, 1985), which was designed to measure two separate domains, student attitudes toward the course they were enrolled in, and student attitudes toward the usefulness of statistics in their field of study. The Statistical Anxiety Rating Scale (Cruise et al., 1985) was designed to measure the value of statistics, the interpretation of statistical information, test anxiety, cognitive skills in statistics, fear of approaching the instructor and fear of statistics. Other similar instruments included the Statistics Attitude Survey (Roberts & Bilderback, 1980), and the Statistics Anxiety Inventory (Zeidner, 1991).

However, although a number of instruments that measure attitudes toward statistics already exist, they all differ in content and configuration (Dauphinee, 1993). For example, although some instruments represent attitudes as a construct with six factors, others regard it as a unidimensional construct which hypothesizes that no meaningful domains exist within attitudes (Roberts & Bilderback, 1980). The identification of the factors that form the structure of the student attitudes toward a research methods course may bear important theoretical and practical implications, especially due to the fact that this has never been examined before. For example, by identifying these subscales of attitudes, research methods instructors may include themselves in the process of learning research from a different angle. By using these domains, instructors may facilitate the learning of research for their students, by enabling them to create more positive attitudes toward such courses. Therefore, the central aims of this study are to explore the multidimensional factor structure of the "Attitudes Toward Research" scale (ATR) and to examine its psychometric properties. This questionnaire was developed by the author of this paper in the Fall of 2002, and the original version consisted of 56 items that were created on a Likert type scale. Based on the analysis of the psychometric properties of this questionnaire that is presented in this paper, further refinements of the questionnaire have been completed, and are presented in section 3. Although the questionnaire was administered in Greek, a translated version of the questionnaire is presented in English in the Appendix.

2. METHOD

2.1. SAMPLE

The data for this study were collected from students who had completed a compulsory and introductory undergraduate course in 'Methodology of Educational Research' at the University of Cyprus. All the students in the sample were enrolled in the elementary or kindergarten education major. This major is considered of very high esteem in Cyprus, and only the highest ability students are accepted into this major. Students from no other majors were obtained from the University since research methods courses are only required for students in the elementary and kindergarten education major. The target population for this study included all students who had completed this course in a period of three years. This population would have consisted of about 450 students. Among the 226 students who took part in the study, 98 (43.4%) completed the questionnaire on the last day of their research methods course, while the remaining 56.6% also answered the questionnaire on the last day of the semester, although they had completed the course one to four semesters earlier. Of the total 226 students in the sample, 15.6% were male and the remaining 84.4% were female. Although there was a disproportionate number of females in the study, this was because there are generally more female than male students that choose to major in elementary or kindergarten education in Cyprus, and the breakdown is not dissimilar to that usually holding in the group majoring in these subjects.

Of the complete sample, 36.9% were sophomores, 34.2% were juniors and 28.9% were seniors. All students who had attended the classes from which the data were collected, responded to the questionnaire, and no non-responses were encountered.

The students were also asked to indicate their self reported level of socioeconomic status (SES), as well as the overall level of their parents' education. Both questions were closed option questions, where the students had to select among four options (very high, high, average and low). In terms of SES, only one student indicated that their level of SES was very high. There were 84.4% of the students that indicated that their SES level was average, 12.8% who indicated that their SES was high, and 2.2% that indicated that their SES was low. In terms of the parents' level of education, 5.3% indicated that their parents' education was very high. About 25% indicated that their parents' level of education was high, while 55.8% considered their parents' level of education to be average. In terms of the Grade Point Average (GPA) that the students had at the University, the majority of the students (57.1%) indicated that their grades ranged from 7.01 to 8.00 points, out of a total of 10 points. There were 15.25% of the students who had grades that ranged between 6.01 to 7.00, while the rest of the students had grades higher than 8.01 points. What is also interesting is that the students were also asked about their high school grade point average (GPA). The results of the study showed that 50% of the students had grades that ranged from 19.01 to 20, out of a total of 20 points, while 34.8% had grades ranging from 18.01 to 19.00. The rest of the students had grades lower than 18.01. The lowest grades were obtained by only a single student who responded as having earned a high school GPA between 12.00. -14.00 out of 20. In addition, 5 students had high school GPAs between 14.01-16.00.

The research methods course in which the students were enrolled, was designed to prepare students to undertake a research project related to educational issues. This course covers the fundamental concepts of research methodology, as well as basic statistical terms and techniques required to analyze research data. Primary emphasis is placed on the research stages; those of conceptualizing and defining a research problem, conducting literature reviews, data collection and analysis techniques, as well as writing and interpreting results, discussions and conclusions in research articles. This course also places substantial emphasis on measurement issues such as scales of measurement, and reliability and validity issues. Finally, the students in this course are required to design and execute a research project related to educational issues throughout the semester.

2.2. STATISTICAL PROCEDURES

The Attitudes Toward Research (ATR) scale that was created by the authors of this paper, consisted of items listed on a 7-point Likert scale. The score 1 represented the option "strongly disagree" while option 7 on the scale represented the category "strongly agree". An initial pool of 56 Likert-type attitudinal items regarding attitudes toward research was constructed. Some items were positively worded and some negatively worded. For the analysis of the data, all negatively worded items were reversed so that a higher numbered response on the Likert scale would represent positive attitudes.

At a preliminary examination, the 56 items of the ATR measure underwent an initial reliability analysis to determine the internal consistency of the items (Andrews & Hatch, 1999). In addition, the product-moment coefficient r between each item and the total score was also calculated. Items which were not significantly related to the total score, or whose coefficient was less than 0.50 were removed from the questionnaire. Forty-one items remained in the pre-final version of the questionnaire.

A principal factor analysis with varimax rotation was then used to create the factor structure of the 41 questions included in the scale (SPSS, 1998). This analysis was used to "reduce a set of observed variables into a relatively small number of components that account for most of the observed variance" (Marcoulides & Hershberger, 1997, p 164). In order to give each factor a clear and distinct meaning for both theoretical interpretation and practical implication, the orthogonal varimax method of rotation was used to minimize the number of variables that have high loadings on more than one factor. To determine the optimum factor solution, the following criteria were used: (a) computation of the percentage of variance extracted, and (b) interpretability of the factors (Comrey & Lee, 1992). A factor loading with absolute value greater than 0.50 was considered sufficiently high to assume a strong relationship between a variable and a factor. Factor loadings less than 0.50 in absolute value were regarded as insignificant, and the items containing such loadings were removed from the scale. In addition, it was decided that factors with only one or two items, even with loadings greater that 0.50, would be excluded from the final version of the scale. Furthermore, with respect to determining the number of factors, only factors with eigenvalues greater than 1.1 were considered as significant (Rummel, 1970). Finally, the factors that were developed from this study were analyzed further with the use of multidimensional scaling. This was done in order to create a map of the locations of the factors in reference to each other, based on their similarities and dissimilarities.

3. RESULTS

For the purpose of examining the reliability of the ATR measure, Cronbach's alpha coefficient was used to measure the internal consistency of the items in the scale. An initial examination of the entire first version of the questionnaire (all 56 items) produced a reliability coefficient of 0.947 which is very satisfactory. Eleven factors were originally extracted, accounting for 66.4% of the variance. However, based on the restrictions included in the methodology section of this paper, several of the items of the original version of the questionnaire were removed because they were considered as inappropriate. Once the inappropriate items were removed, 32 items remained in the scale. Once the factor analysis was re-run with those items, a five-factor solution remained, which included a robust set of constructs that were relatively easily interpreted. These five factors accounted for 66.25% of the total variance. Details of the items included in the final version of the scale are presented in the Appendix.

The results of the factor analysis have produced a five factor solution. The first factor was clearly the most important one since it accounted for 18.92% of the total ATR scale variance. All items in this factor with loadings greater than 0.50 had to do with the students' opinions about the usefulness of research in their careers. This factor consisted of 9 items, while the two items with the highest loadings on this factor were those of 'research is useful for my career' and 'research is connected to my field of study'. This factor therefore was named 'research usefulness in profession'. This usefulness is interpreted as the perception that students have in terms of how research will be useful and help them in their professional lives.

The second factor accounted for 17.94% of the variance and included items describing tension, stress, fear, difficulties in understanding research, and was called 'research anxiety'. This factor consisted of eight items. The two questions with the highest loadings on this factor were those of 'research makes me nervous' and 'research is stressful.' The third factor which was composed of eight items accounted for 15.42% of the variance and was labeled as 'positive attitudes toward research'. The two questions with the highest loadings on this factor were those of 'I love research' and 'I enjoy research'. The fourth factor accounted for 8.30% of the variance, and consisted of four items referring to the use of research in a student's personal life, and was therefore called 'relevance to life'. The two items with the highest loadings on this factor were those of 'I use research in my daily life' and 'Research oriented thinking plays an important role in everyday life.' The last factor, 'research difficulty', accounted for 5.67% of the total variance. This factor that consisted of only three items included items related to 'having trouble with arithmetic' and 'finding it difficult to understand the concepts of research'. The results of the factor analysis with the loadings of the five factors are presented in Table 1. The items labeled as "Recoded" are listed this way so that all of the items with high values on the Likert scale represent high agreement levels in terms of the respondents' positive attitudes.

The responses on the remaining 32 items on the ATR scale indicated a high reliability for the test, ($r=0.948$). The coefficient alpha reliabilities for the responses to items on each of the five subscales were relatively high. Coefficient alpha reliability for the research usefulness in the profession factor was 0.919 (9 items); for the research anxiety factor it equaled 0.918 (8 items); the reliability for the positive attitudes toward research factor equaled 0.929 (8 items). The reliability of the life relevancy factor equaled 0.767 (4 items), while the reliability for the research difficulty factor equaled 0.711 (3 items).

After the factor analysis was performed, a score was calculated for each student on each factor by obtaining the mean for all items comprising each factor. The mean score of the students on the research usefulness for the profession factor was $F1=5.20$; for the research anxiety factor the mean was 3.17; the mean of the positive attitudes toward research factor was 3.90; for the relevance to life factor the mean score was 5.04, while the mean score of the research difficulty factor was 4.84.

Table 1. Rotated factor loadings of the ATR scale

	Component				
	F1 Research usefulness for profession	F2 Research Anxiety	F3 Positive attitudes toward research	F4 Relevance to life	F5 Research difficulty
Research is useful for my career	.822	.067	.209	.171	.022
Research is connected to my field of study	.783	.107	.282	.037	-.006
Research should be indispensable in my professional training	.768	.087	.264	.272	.132
Research should be taught to all students	.738	.133	.259	.017	.112
Research is useful to every professional	.667	.036	.174	.377	.112
Research is very valuable	.658	.127	.086	.160	.124
I will employ research approaches in my profession	.649	.130	.148	.330	-.029
The skills I have acquired in research will be helpful to me in the future	.608	.164	.296	.418	.051
Knowledge from research is as useful as writing	.601	.087	.285	.377	-.165
Research makes me nervous-RECODED	.156	.857	.189	.080	.077
Research is stressful-RECODED	.197	.807	.239	.054	-.019
Research makes me anxious-RECODED	.220	.798	.217	.010	-.085
Research scares me-RECODED	.160	.794	.155	.024	.161
Research is a complex subject-RECODED	.048	.766	.242	.016	.090
Research is complicated-RECODED	.079	.700	.265	.157	.172
Research is difficult-RECODED	.137	.678	.284	.123	.102
I feel insecure concerning the analysis of research data - RECODED	-.089	.590	.017	.108	.197
I love research	.207	.318	.812	.125	.039
I enjoy research	.222	.268	.789	.077	.034
I like research	.232	.345	.775	.109	.074
I am interested in research	.338	.254	.736	.111	.176
Research acquired knowledge is as useful as arithmetic	.186	.352	.723	.233	.049
Research is interesting	.383	.115	.655	.101	.181
Most students benefit from research	.499	.177	.517	.142	.154
I am inclined to study the details of research	.446	.199	.511	.073	.032
I use research in my daily life	.163	.043	.235	.752	-.008
Research-orientated thinking plays an important role in everyday life	.391	.040	.086	.688	.060
Research thinking does not apply to my personal life-RECODED	.398	.210	-.046	.598	.144
Research is irrelevant to my life-RECODED	.408	.163	.200	.569	.081
I have trouble with arithmetic-RECODED	.074	.060	.137	.012	.792
I find it difficult to understand the concepts of research-RECODED	.146	.427	.062	.204	.686
I make many mistakes in research-RECODED	.096	.518	.203	.005	.610

A test developed by Hotelling, called Hotelling's T^2, was then applied to the data. This test allows for the comparison of several observed means, five in our case, to a set of constants, which was the median of the seven point Likert scale that was used in the ATR measure. The results of the

MANOVA indicated statistical significance (Hotelling's T^2 =30.967, $p<0.01$). Since the hypothesis of no differences was rejected, the univariate test was used to get an idea of where the difference among each of the five subscales compared to the median of 4 may lie. The results are summarized in Table 2. Thus, as a group, students consider research to be useful in their professional lives, and in their personal lives (relevance to life). However, the students tended to have quite negative attitudes toward research as well as anxiety toward the subject, although they responded that they did not have a lot of difficulty in understanding this subject. The factor that deviated the most from the median was research usefulness, indicating that the students truly understood and appreciated the usefulness of research in their professional lives. The next highest factor was that of relevancy of research in the student's personal lives. The factor that deviated the least from the median was that of positive attitudes toward research. This indicated that although the students indicated that they had some negative attitudes toward this subject, they did not deviate a lot from the median indicating that their responses were actually quite neutral in terms of attitudes. The overall students' attitudes toward research, when taking into account all seven dimensions is 4.43 which is positive although it is actually closer to the median of the seven point Likert scale.

Correlation coefficients between the Attitudes Toward Research sub-scales were also calculated. As presented in Table 3, the intercorrelations of the Attitudes Toward Research factors suggested the following pattern of interrelationships. The research usefulness factor was most highly correlated with the factors of relevancy to life (r=0.69) and with the factor of positive attitudes toward research (r=0.67). The anxiety subscale was most highly correlated with the positive attitudes (r=0.58) and the difficulty (r=0.52) factors. Finally, the research difficulty factor was most highly correlated with the research anxiety factor (r=0.52).

Table 2. Cell means, standard deviations and univariate F-tests

Factors	X	s	Hypoth. SS	Error MS	F	p
F1 research usefulness	5.19	1.13	5870.017	1.28	4582.07	0.000
F2 research anxiety	3.19	1.27	2222.091	1.62	1368.89	0.000
F3 positive attitudes	3.91	1.23	3337.655	1.50	2223.60	0.000
F4 relevance to life	5.04	1.10	5530.293	1.21	4559.43	0.000
F5 difficulty of research	4.84	1.21	5102.393	1.63	3129.25	0.000

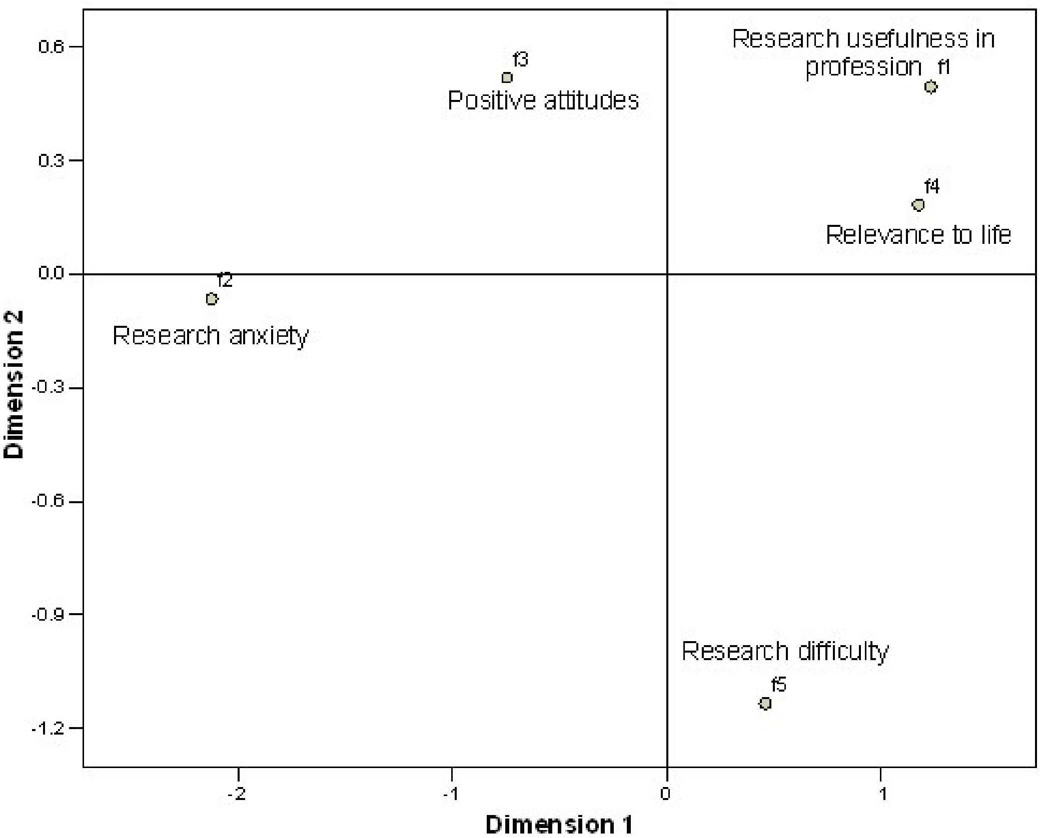
Table 3. Inter-correlations between the five factors

	Research usefulness for profession	Research anxiety	Positive attitudes toward research	Relevance to life
Research anxiety	0.363(**)			
Positive attitudes toward research	0.671(**)	0.587(**)		
Relevance to life	0.697(**)	0.324(**)	0.485(**)	
Research difficulty	0.290(**)	0.527(**)	0.424(**)	0.287(**)

**p<0.01 for all correlations

There are many different measures for quantifying similarity, and the Pearson correlation coefficient is one of those most frequently used (Norusis, 1990). However, similarity measures can also estimate the degree of closeness between objects. For this study, a multidimensional scaling analysis was performed with the statistical package SPSS. Multidimensional scaling was used in order to be able to display multivariate data (the five factors in this case) on a lower two-dimensional space. This is done by mapping the distances between points in a high dimensional space into a lower dimensional space (Johnson, 1998). Four clusters have resulted from the use of these measures (see Figure 1). Projecting the points of each of the factors on Dimension 1 of the axis reveals two different groups of factors: the first group is comprised of the factors dealing with the usefulness of research in the student's professional and personal lives and goes along with the factor of research difficulty. The second group in the same dimension contains factors that deal with attitudinal issues related to the subject of research methods (positive attitudes and research anxiety). However, by projecting the points on Dimension 2, two different groups are created. Group 1 includes the factors of research anxiety, and research difficulty, while group 2 includes the factors of research usefulness (in the students' personal and professional lives) and positive attitudes toward research. This distinction again shows that on the one hand research anxiety and difficulty seem to interrelate, while the positive attitudes toward research appear to group together with the usefulness of research. Overall however, by looking at the two dimensions it is clear that the usefulness of research factors are constantly grouped together, and are never grouped together with the research anxiety factor. This is to show that research anxiety could possibly stem from other factors that have nothing to do with whether the students consider research to be useful in their lives or not. In addition, positive attitudes toward research are never grouped together with the factor of positive attitudes toward research. This again shows that there are different factors that can possibly influence the student's attitudes toward this subject, that have nothing to do with whether they consider a research methods course to be difficult or not.

Figure 1. Two dimensional configuration of the five factor model based on Euclidian distances



4. DISCUSSION

The major objective of this study was to verify the domains of attitudes related to research among education undergraduate students. The majority of the instruments designed to measure attitudes, have been focused on statistics, and have produced configurations of attitudes ranging from one to six dimensions. Although there may be some degree of similarity in the attitudes between statistics courses and research methods courses, none of the instruments related specifically to attitudes toward research. One definition representing a configuration of attitudes toward research was created by the

Attitudes Toward Research (ATR) measure. The current study based on the ATR measure indicated that students' attitudes toward research are comprised of seven areas.

More specifically, an exploratory factor analysis using undergraduate students indicated that the ATR measure consists of five meaningful factors. The first factor is that of the usefulness of research in the student's professional life. The second factor is that of research anxiety. The third factor is that of positive attitudes toward research. The fourth factor is that of relevancy to the student's non-academic and non-professional lives, which is comprised of attitudes about the use of research in the student's life, while the fifth factor is that of the difficulty of research.

This study has also examined the relationships that existed between the five factors that were produced in this study. Overall, the strongest relationship existed between the usefulness factor and the relevancy to life factors. This confirms a common observation about human attitudes: people feel favorably toward activities, or objects that are useful in their lives. Another strong relationship that was found in the data had to do with affective factors, including those of research anxiety, research difficulty and positive attitudes toward research. These results indicate that there are basically two main groups of factors that are influencing the study's results. On the one hand students tend to form some affective views toward research, that may or may not be influenced by whether they consider research to be a useful subject or not. More specifically, although the usefulness of research for the profession and in daily life is highly correlated with the positive attitudes factor, this is not the case with the factors of research difficulty and anxiety. This indicates that students who can see the usefulness of research also tend to have more positive attitudes toward the subject. However, issues of whether research is difficult, or if it causes anxiety to the students do not appear to be highly correlated with the usefulness factors.

By identifying the five factors that comprise students' attitudes toward research, instructors may begin discussions about the importance of learning research and its importance on making academic and professional career choices. In addition, by using information from these domain areas, instructors may be able to identify specific modifications to attitudes, skills and behaviors to facilitate the learning of research and foster a deeper appreciation of this subject. The availability of an instrument such as the ATR scale which has been designed for students, may provide information concerning motivational aspects associated with learning research, and might also have potential for identifying distinctive attitude profiles of students who find research problematic. Overall however, this study's results validate the utility of the ATR scale in measuring student attitudes toward research.

The results of this study also need to be re-examined to determine if they can be replicated with other samples of students, as well as with different populations. In addition, the future exploration of the relationships between attitudes and student achievement in research is an important area that still needs to be examined further. Finally, it would also be useful to examine the process of attitude change of students, and what it is based on, by collecting student data at various points in the semester. With the use of structural equation modeling, these variables could all be integrated in a single analysis to determine how these variables all influence each other.

REFERENCES

Adams, N. A., & Holcomb, W. R. (1986). Analysis of the relationship between anxiety about mathematics and performance. *Psychological Reports, 59*, 943-948.

Andrews, P., & Hatch, G. (1999). A new look at secondary teachers' conceptions of mathematics and its teaching. *British Educational Research Journal, 25*(2), 203-223.

Comrey, L., & Lee, H. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *American Statistical Association Proceedings of the Section on Statistics Education,* 92-97.

Dauphinee, T. L. (1993). *The factor structure and gender factor invariance of an instrument to assess adult attitudes toward statistics.* Unpublished doctoral dissertation, The University of New Mexico, Albuquerque, New Mexico.

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

Elmore, P. B., & Vasu, E. S. (1980). Relationship between selected variables and statistics achievement: Building a theoretical model. *Journal of Educational Psychology, 46*, 215-222.

Elmore, P. B., & Lewis, E. L. (1991, April). *Statistics and computer attitudes and achievement of students enrolled in applied statistics: Effect of computer laboratory.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Johnson, D. E. (1998). *Applied multivariate methods for data analysis*. Pacific Grove, CA: Duxbury Press.

Ma, X. (1995). Factor structure of attitudes toward mathematics among high, average, and low achieving high school senior students in the Dominican Republic. *Focus on Learning Problems in Mathematics, 17*(4), 20-35.

Marcoulides, G. A., & Hershberger, S. L. (1997*). Multivariate statistical methods: A first course*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*, 60-70.

Norusis, M. J. (1990). *SPSS: Statistical data analysis*. Chicago: SPSS Inc.

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and validity of a statistics attitude survey. *Educational and Psychological Measurement, 40*, 235-238.

Rummel, R. J. (1970). *Applied factor analysis.* Evanston: Northwestern University Press.

Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. D. (1995*).* The development and validation of attitudes towards statistics. *Educational and Psychological Measurement, 55*(5), 868-875.

SPSS, Inc. (1998). *SPSS Base 8.0: Applications guide.* Chicago, IL: Author.

Waters, L. K., Martelli, T. A., Zakrajsek, T., & Popovich, P. M. (1988). Attitudes toward statistics: An evaluation of multiple measures. *Educational and Psychological Methods, 48*, 513-516.

Wise, S. L. (1985) The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement, 45*, 401-405.

Woelke, P. L. (1991, April). *An examination of the factor structure of Wise's attitude toward statistics scale*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology, 61*, 319-328.

ELENA C. PAPANASTASIOU
Intercollege, Cyprus
Mourouzi 7, #402
1055 Nicosia
Cyprus

APPENDIX

STUDENTS' "ATTITUDES TOWARD RESEARCH" SCALE **

The following statements refer to some aspects of educational research. Please answer all the questions sincerely. DO NOT DISCLOSE YOUR IDENTITY ANYWHERE.

Circle one of the numbers opposite each of the statements that follow.

By selecting number 1 you indicate that you strongly disagree.

By selecting number 7 you indicate that you strongly agree.

	Strongly Disagree						Strongly Agree
1. Research makes me anxious *	1	2	3	4	5	6	7
2. Research should be taught to all students	1	2	3	4	5	6	7
3. I enjoy research	1	2	3	4	5	6	7
4. Research is interesting	1	2	3	4	5	6	7
5. I like research	1	2	3	4	5	6	7
6. I feel insecure concerning the analysis of research data *	1	2	3	4	5	6	7
7. Research scares me *	1	2	3	4	5	6	7
8. Research is useful for my career	1	2	3	4	5	6	7
9. I find it difficult to understand the concepts of research *	1	2	3	4	5	6	7
10. I make many mistakes in research *	1	2	3	4	5	6	7
11. I have trouble with arithmetic *	1	2	3	4	5	6	7
12. I love research	1	2	3	4	5	6	7
13. I am interested in research	1	2	3	4	5	6	7
14. Research is connected to my field of study	1	2	3	4	5	6	7
15. Most students benefit from research	1	2	3	4	5	6	7
16. Research is stressful *	1	2	3	4	5	6	7
17. Research is very valuable	1	2	3	4	5	6	7
18. Research makes me nervous *	1	2	3	4	5	6	7
19. I use research in my daily life	1	2	3	4	5	6	7
20. The skills I have acquired in research will be helpful to me in the future	1	2	3	4	5	6	7
21. Research is useful to every professional	1	2	3	4	5	6	7
22. Knowledge from research is as useful as writing	1	2	3	4	5	6	7
23. Research is irrelevant to my life *	1	2	3	4	5	6	7
24. Research should be indispensable in my professional training	1	2	3	4	5	6	7
25. Research is complicated *	1	2	3	4	5	6	7
26. Research thinking does not apply to my personal life *	1	2	3	4	5	6	7
27. I will employ research approaches in my profession	1	2	3	4	5	6	7
28. Research is difficult *	1	2	3	4	5	6	7
29. I am inclined to study the details of research procedures carefully	1	2	3	4	5	6	7

30. Research is pleasant	1	2	3	4	5	6	7
31. Research-orientated thinking plays an important role in my daily life	1	2	3	4	5	6	7
32. Research is a complex subject *	1	2	3	4	5	6	7

* The items with an asterisk are items whose direction has been changed in the analysis.

** This version of the questionnaire has been translated to English from Greek.

"VARIATION-TALK": ARTICULATING MEANING IN STATISTICS

KATIE MAKAR
University of Queensland, Australia
k.makar@uq.edu.au

JERE CONFREY
Washington University in St. Louis, U.S.A.
jconfrey@wustl.edu

SUMMARY

Little is known about the way that teachers articulate notions of variation in their own words. The study reported here was conducted with 17 prospective secondary math and science teachers enrolled in a preservice teacher education course which engaged them in statistical inquiry of testing data. This qualitative study examines how these preservice teachers articulated notions of variation as they compared two distributions. Although the teachers made use of standard statistical language, they also expressed rich views of variation through nonstandard terminology. This paper details the statistical language used by the prospective teachers, categorizing both standard and nonstandard expressions. Their nonstandard language revealed strong relationships between expressions of variation and expressions of distribution. Implications and the benefits of nonstandard language in statistics are outlined.

Keywords: Statistical reasoning; Statistics education; Reasoning about variation and distribution; Mathematics education; Teacher education; Nonstandard language

Do not let calculations override common sense.
Moore & McCabe (1993, p. 121)

1. INTRODUCTION

Recently, researchers in statistics education have been calling for greater emphasis in schools on developing students' conceptions of variation (Moore, 1990; Shaughnessy, Watson, Moritz, & Reading, 1999; Pfannkuch & Begg, 2004). In addition, they argue that too much instruction in statistics has focused on performing statistical operations rather than developing students' thinking about what makes *sense*. One approach to sense-making is through encouraging learners to express their ideas in their own words (Noss & Hoyles, 1996). Russell and Mokros (1990) documented teachers' statistical thinking by attending to their ways of describing data using nonstandard statistical language. Other research has also revealed that learners often articulate concepts of variation using nonstandard language (e.g., Bakker, 2004; Reading, 2004); however these studies have not systematically looked at ways that learners express notions of variation. Even less is known about how teachers express concepts of variation.

In other qualitative studies, we examined preservice secondary mathematics and science teachers' experiences making sense of data as they conducted technology-based investigations (Confrey, Makar, & Kazak, 2004; Makar, 2004; Makar & Confrey, submitted). This paper reports on an exploratory study of preservice teachers' use of standard and nonstandard statistical language when discussing notions of variation. Below we first review the background for the present study, including

Statistics Education Research Journal, 4(1), 27-54, http://www.stat.auckland.ac.nz/serj

the conceptualization of variation, prior research on teachers' and students' knowledge of variation, and research needs regarding meaning construction and language use in statistics education.

2. REASONING ABOUT VARIATION

Although variation is a central component of understanding statistics, little is known about how children (and even less about how teachers) reason with, conceptualize, learn about, and *express* notions of variation. This section begins with a conceptual analysis of variation, then looks at recent research on developing teachers' conceptions of variation, and finally turns to recent work on middle school students' reasoning in this area.

2.1. VARIATION AS A CONCEPTUAL ENTITY

Most uses of the term "variation" in research studies are taken to have a self-evident, common sense meaning, and leave it undefined. Variation is closely linked to the concepts of variable and uncertainty. It is often regarded as a measurement of the amount that data deviate from a measure of center, such as with the interquartile range or standard deviation. Variation encompasses more than a measure, although measuring variation is an important component in data analysis. In considering variation, one must consider not just *what* it is (its definition or formula), or *how* to use it as a tool (related procedures), but also *why* it is useful within a context (purpose).

In simple terms, variation is the quality of an entity (a variable) to vary, including variation due to uncertainty. "Uncertainty and variability are closely related: because there is variability, we live in uncertainty, and because not everything is determined or certain, there is variability" (Bakker, 2004, p. 14). While no one expects all five year-old children to be the same height, there is often difficulty understanding the extent to which children's height vary, and that the variability of their heights is a mixture of explained factors (e.g., the parents' heights, nutrition) and chance or unexplained factors.

In this paper, we will not distinguish between reasoning about variability and reasoning about variation, although Reading and Shaughnessy's (2004, p. 202) definition of variation distinguishes it from variability:

> The term *variability* will be taken to mean the [varying] characteristic of the entity that is observable, and the term *variation* to mean the describing or measuring of that characteristic. Consequently, the following discourse, relating to "reasoning about variation," will deal with the cognitive processes involved in describing the observed phenomenon in situations that exhibit variability, or the propensity for change.

We would argue that developing the concept of variation, beyond just an acknowledgement of its existence, requires some understanding of distribution. In examining a variable in a data set, one is often interested in uncovering patterns in the variability of the data (Bakker, 2004), for example by representing the ordered values of the variable in a graph. The distribution, then, becomes a visual representation of the data's variation and understanding learners' concept of variation becomes closely linked to understanding their concept of distribution. A key goal in developing students' reasoning about distributions is assisting them in seeing a distribution as an *aggregate* with its own characteristics (such as its shape or its mean) rather than thinking of a distribution as a collection of individual points (Hancock, Kaput, & Goldsmith, 1992; Konold & Higgins, 2002; Marshall, Makar, & Kazak, 2002; Bakker, 2004).

In thinking about what the notion of distribution entails, one might think of the associated measures, properties, or characteristics of a distribution—for example, its mean, shape, outliers, or standard deviation. These entities in isolation, however, do not capture the desired concept and can aggravate the focus on individual points. In addition, they lack the idea that we want to capture a distribution *of* something. Pfannkuch, Budgett, Parsonage, and Horring (2004) cautioned that students often focus on characteristics of a distribution, but forget to focus on the meaning of the distribution within the context of the problem. In comparing distributions, for example, they suggested that students should first look at the distributions and compare centers, spread, and anything noteworthy, but then should be asked to draw an evidence-based conclusion, using probabilistic rather than deterministic language, based on their observations.

2.2. TEACHERS' REASONING ABOUT VARIATION

Teachers' mind-set and conceptions about data have been of great concern in the research community for over a decade (Hawkins, 1990; Shaughnessy, 1992). Despite this, Shaughnessy (2001) reports, "[I am] not aware of any research studies that have dealt specifically with teachers' conceptions of variability, although in our work teaching statistics courses for middle school and secondary school mathematics teachers we have evidence that many teachers have a knowledge gap about the role and importance of variability."

Three very recent research projects have focused specifically on teachers' notions of variation and distribution. Hammerman and Rubin (2004) report on a study of a two-year professional development project that gave secondary teachers an opportunity to analyze data with a new data visualization tool. The software Tinkerplots™ (Konold & Miller, 2004) allows learners to create their own data representations through sorting, ordering, and stacking data. They found that teachers did not choose to use measures of center in comparing distributions even though those measures were readily available in the software, but rather chose to compare groups by comparing slices and subsets of distributions to support compelling arguments with data. In addition, Hammerman and Rubin noted that teachers were dissatisfied with representations such as histograms and box plots that would hide the detail of the underlying distribution. From their study, we can begin to see the potential that new visualization tools have for helping teachers to construct their own meaning of statistical concepts.

In another study, Canada (2004) created a framework to investigate prospective elementary teachers' expectation, display, and interpretation of variation within three statistical contexts: repeated sampling, data distributions, and probability outcomes. His study took place in a preservice mathematics course developed to build the teachers' content knowledge of probability and statistics through activities that emphasized hands-on experiments involving chance and computer-generated data simulations. Canada's study found that initially, although the prospective teachers had a good sense of center, they found it difficult to predict realistic spreads in distributions from collected data or produced from random sampling and probability experiments. Their predictions for underlying distributions included expectations of way too little variation, overly extreme variation, or unrealistically symmetrical distributions. In addition, many of the teachers took an initial stance that 'anything goes' in random experiments and felt that they could therefore make no judgment about the distribution of outcomes. After the course, the teachers demonstrated stronger intuitions about variation - their predictions were much more realistic and their expectation of variation more balanced. In addition, the teachers' descriptions of distributions were more rich and robust in their recognition of variation and distribution as an important concept, for example by referencing how the distributions were clustered, spread, concentrated, or distributed. Canada's study hints at a potential link between the elements that teachers focus on when describing distributions and their intuition about variation and uncertainty in data.

In previous research on statistical inquiry with practicing middle school teachers, we found that teachers, like their students, often begin examining data distributions by focusing on individual points (Confrey & Makar, 2002). Yet in a meaningful context (e.g., student test scores), the teachers in our study constructed a need for examining variation in a distribution by initiating a discussion of how the distribution of students' abilities affected their choice of instructional strategies. We noticed, however, that when the teachers did not adequately construct more complex concepts, like sampling distributions, they would tend to use statistical tools mechanically, without carefully examining their relationship to the data (Makar & Confrey, 2004).

In summary, although much concern has been expressed about the over-emphasis on procedures in teaching statistics, the studies described above highlight some important ways to improve what Shaughnessy (1992) calls teachers' lack of intuition about stochastics. The main ideas these studies bring out is that teachers themselves need to learn statistical concepts in an environment much like the one recommended for students – one that is active, technology-rich, involving authentic data, and offering plenty of opportunities to build their conceptions through experiences with data. Particulars about how teachers build these conceptions still needs further research.

2.3. STUDENTS' REASONING ABOUT VARIATION

It is useful to briefly review research on students' development of concepts of variation, as it may provide insight into potential ways to build teachers' conceptions of variation. First, teachers frequently possess similar reasoning to their students (Hammerman & Rubin, 2004; McClain, 2002). Second, understanding students' conceptions of variation can help teachers to plan instruction. Finally, the research in this area has uncovered new insights into students' intuitions about variation without formal procedures and terminology. These insights may provide ideas for new methods of professional development for teachers.

A common thread in the research on teaching concepts of variation and distribution is a recommendation to focus not only on the characteristics of distributions, but on their *purpose*. A well-documented approach to this is through comparing groups (see for example, Cobb, 1999; Watson & Moritz, 1999; Biehler, 1997; Makar & Confrey, 2004; Hammerman & Rubin, 2004). Pfannkuch et al. (2004) found that the descriptions of distributions given by secondary students sometimes included characteristics of the distribution that were disconnected from the context or meaning of the problem. For example, when taking note of the variability of the data, half of the students compared the ranges, which was not relevant to their question, and most of the students focused on comparing measures of center or extremes. Pfannkuch and her colleagues hypothesized that the instruction the students received focused not on drawing meaningful conclusions, but rather on comparing features of the box plot. They recommended that instruction concentrate not on *how* to compare centers, but *why* one should do so. Reading and Shaughnessy (2004) also found that tasks which asked students not just for descriptive information but also for *explanations* offered greater possibilities for insight into students' reasoning about variation.

In developing students' conception of statistics, Konold and Pollatsek (2002) argue that "the central idea should be that of searching for a signal and that the idea of distribution comes into better focus when it is viewed as the 'distribution around' a signal" (p. 262). Bakker and Gravemeijer (2004) on the other hand hold that "reasoning with shapes forms the basis for reasoning about distributions" (p. 149). By developing a lens of seeing the distribution as an entity, one can then look at statistical measures as characteristics of the distribution rather than as calculations from individual points (Bakker, 2004).

Bakker (2004; Bakker & Gravemeijer, 2004) also found that with Cobb's Minitool software and an innovative learning trajectory he was able to encourage his middle school students to think about variation and towards a distribution-view of data. Initial discussions with students focused on the middle "bump" of the distribution, but later the "bump" came to represent the whole distribution. Konold and his colleagues (2002), who studied middle school students, emphasized that in problem solving with data, the middle bump was a frequently identified portion of a mound-shaped distribution, and termed this central bump a *modal clump*. Bakker (2004) also found that rather than just focusing on the central region of a distribution, students also tend to divide distributions into three categories: low, middle, and high. Hammerman and Rubin (2004) saw the teachers in their study take a similar tactic in comparing distributions. These findings may indicate that dividing distributions into three pieces may be a more natural way for students to examine a distribution than by dividing it into four sections, as in the box plot.

The studies reported above suggest that if learners are provided with relevant contexts, concrete experiences, complex tasks, adequate time and support, and appropriate tools to build statistical concepts, then their understanding appears to be much more robust. One notable common feature of these studies is that they developed learners' notions of variation through constructing a purpose for variation, often building on the learners' own language for describing and interpreting what they were seeing.

3. CONCEPT DEVELOPMENT

Below we briefly discuss some aspects of how learners construct meaning of data as well as issues surrounding the use of learners' everyday or nontechnical terminology. These issues affected the design of the present study and the lens we used to interpret prospective teachers' articulation of variation. The need to refer both to standard statistical terminology and to nonstandard or nontechnical statistical terminology arises because one cannot assume that if a respondent can repeat the definition of a concept, she has necessarily assimilated this concept. For example, although a student may be taught the concept and formula for standard variation and can even use this term as part of class discourse, this does not imply that they are "seeing" variation in what is being measured.

3.1. CONSTRUCTING MEANING

Because students are frequently taught definitions and procedures without first developing their own intuition and meaning about the concepts underlying them, premature instruction of formal terminology and rules can inhibit students' own sense-making (Flyvbjerg, 2001; Schoenfeld, 1991; Boaler, 1997). Formal mathematical procedures, terminology, and symbolism are critical for developing advanced levels of mathematical understanding in that they can provide efficient paths to problem-solving, focus attention on particular aspects of a problem, and open new levels of understanding of the concepts represented by the terms or symbols. However, the emphasis must be on building meaning, not simply assuming that standard procedures or terms can themselves carry the meanings of underlying concepts.

Shaughnessy (1992) notes that students' and teachers' lack of intuition about stochastics is a critical barrier to improved teaching and learning in statistics. Fischbein's (1987) description of intuition matches well with the kind of thinking that we believed should be developed in teachers.

> *Intuitions are always the product of personal experience, of the personal involvement of the individual in a certain practical or theoretical activity.* Only when striving to cope actively with certain situations does one need such global, anticipatory, apparently self-consistent representations. The development of … intuitions implies, then, didactical situations in which the student is asked to evaluate, to conjecture, to predict, to devise, and check solutions. In order to develop new, correct probabilistic intuitions, for instance, it is necessary to create situations in which the student has to cope, *practically*, with uncertain events (p. 213, italics in original).

For building intuition, Fischbein further argues that the role of visualization "is so important that very often intuitive knowledge is identified with visual representations" (p. 103). Insight can also be gained into learners' sense-making by focusing on how they use their own words to explain relationships, constructs, and processes (Noss & Hoyles, 1996). One might infer from these last three quotes that intuition in statistics is built through the development of visualization, articulation, and representation of data distributions within personally compelling contexts. Intuition about variation, then, may be fostered through a lens of "seeing variation" (Watkins, Scheaffer, & Cobb, 2003, p. 10), or Moore's (1990) acknowledgement of the omnipresence of variation. These insights may be gained through attending to students' own words for describing what they are seeing. These words are often less technical and contain elements of nonstandard language.

3.2. NONSTANDARD LANGUAGE USE

Noss and Hoyles (1996) state that focusing on mathematical meaning moves us to consider how learners *express* mathematics rather than how they *learn* it. The difference is in the locus of the concept. If we focus on how students learn statistics, it implies that statistical concepts are ontologically fixed and that the goal of learning is to impart *a priori* knowledge from teacher to student. However, by turning that around to foster learners' dynamic conceptions of statistical concepts, we acknowledge that these concepts are not passively received, but rather are actively and socially constructed by the individual.

Developing understanding oftentimes requires the use of nonstandard terminology. Unfortunately, when students do use their own language to make meaning, teachers often do not recognize nonstandard ways of talking (Lemke, 1990). Biehler (1997) attributes some of the difficulty in communicating and understanding relationships in data to the lack of formal language we have in describing distributions beyond the level of statistical summaries:

> The description and interpretation of statistical graphs and other results is also a difficult problem for interviewers and teachers. We must be more careful in developing a language for this purpose and becoming aware of the difficulties inherent in relating different systems of representation. Often, diagrams involve expressing relations of relations between numbers. An adequate verbalization is difficult to achieve and the precise wording of it is often critical. There are profound problems to overcome in interpreting and verbally describing statistical graphs and tables that are related to the limited expressability of complex quantitative relations by means of common language (p. 176).

Biehler's work implies that the current focus on statistical summaries in describing distributions is inadequate and that research needs to be improved in the area of developing statistical language for interpreting more robust relationships in data.

4. DESIGN AND METHODOLOGY

4.1. APPROACH

This exploratory study was designed to gain insight into the ways that prospective secondary mathematics and science teachers express or discuss notions of variation when engaged in a purposeful statistical task. Based on the literature reviewed, it was anticipated that respondents will use standard or conventional descriptions and terms, as well as nonstandard descriptions and informal language, and the goal of the study was to document the different types of language used. Respondents were interviewed twice using an identical task, in the first and last week of a fifteen-week preservice course on assessment, which included an embedded component of exploratory data analysis. The task given during the interview asked teachers to compare two distributions of data relevant to the context of teaching, in terms of the relative improvement in test scores of two groups of students.

Although interviews were conducted at the beginning and end of the course, the purpose of the study was not to evaluate the effectiveness of the course or to compare performance before and after instruction. Some comparisons will be made of language use before and after the course, but given the small number of respondents, these comparisons should be interpreted with caution. The primary interest was in categorizing and describing the language respondents use to describe variation in the data. Any change that may have occurred in respondents' language use simply enriched and extended the range of responses available for analysis.

4.2. SUBJECTS

The respondents for the study were secondary mathematics and science preservice teachers at a large university in the southern United States enrolled in an innovative one-semester undergraduate course on assessment designed and taught by the authors (Confrey et al., 2004). Twenty-two students began the course, but four withdrew from the course before the end of the semester. In addition, some data from one subject was lost due to technical malfunction, leaving seventeen subjects with a complete set of data. The seventeen subjects ranged in age from 19 to 42, with ten students of traditional college age (19-22), five aged 23-29, and two students over 30 years of age. Of these, three were male and fourteen female, nine were mathematics majors and eight were science majors (predominantly biology). About 60% of the class was Anglo, with the remaining students being of Hispanic and African-American ancestry. The students had varying backgrounds in statistics: eight had not previously studied statistics, while five had previously taken a traditional university-based statistics course either in the mathematics department or in a social science department. The remaining four had not taken a formal course, but had previous experience in statistics as a topic in one of their mathematics or science courses.

4.3. SETTING

The study took place at the beginning and end of a one-semester course that integrated ideas of assessment, data analysis, equity, and inquiry—themes identified as critical but missing from preservice education (National Research Council, 1999; 2000; 2001). The purpose of the course was to give the preservice teachers some background in classroom and high-stakes assessment, develop their statistical reasoning, gain experience using technological tools to interpret student assessment results, and to introduce them to issues of equity through examining data. In the final month of the course, the prospective teachers conducted their own data-based inquiry into an issue of equity in assessment.

The prospective teachers were guided through several investigations that built an atmosphere of interpretation of data rather than the development of formal theoretical foundations. The inclusion of the dynamic statistical software Fathom™ (Finzer, 2001) was critical as a learning tool rather than a traditional statistical package aimed at statisticians and statistics students. The statistical content of the course was comprised of an overview of data graphing (histograms, box plots, dot plots), descriptive statistics (mean, median, standard deviation, interquartile range, distribution shapes), linear regression (association, correlation, least-squares, residuals), and a brief introduction to sampling distributions and inference (through building of simulations). The statistical content was developed as a set of tools to gain insight into data rather than as isolated topics.

4.4. INTERVIEW TASK AND PROCEDURE

Respondents were interviewed by the first author during the first and last week of the course, using the same task in both interviews. Most interviews lasted between ten and twenty minutes. The task was set in the context of an urban middle school trying to determine the effectiveness of a semester-long mathematics remediation program (called 'Enrichment') for eighth-grade (13-14 years old) students considered in need of extra help preparing for the state exam given each spring. To decide if the Enrichment program was working, the school compared student scores from their seventh grade state exam score in mathematics to their scores on a practice test given near the end of eighth grade. Respondents were shown a pair of dot plots (Figure 1) of authentic data taken from students in an Enrichment class (upper distribution) and a regular eighth grade class (lower distribution). Respondents were asked initially to compare the relative improvement of students in the two groups, and then were probed if their responses needed clarification. It is important to note that the data in Figure 1 represent the *change* (difference) in scores between the two assessments, i.e. numbers on the x-axis are positive when scores improved, and negative when scores declined. (Data points in red (shaded dots) highlighted those students classified as economically disadvantaged; this was used for another question in the interview related to equity but not pertinent for this study.)

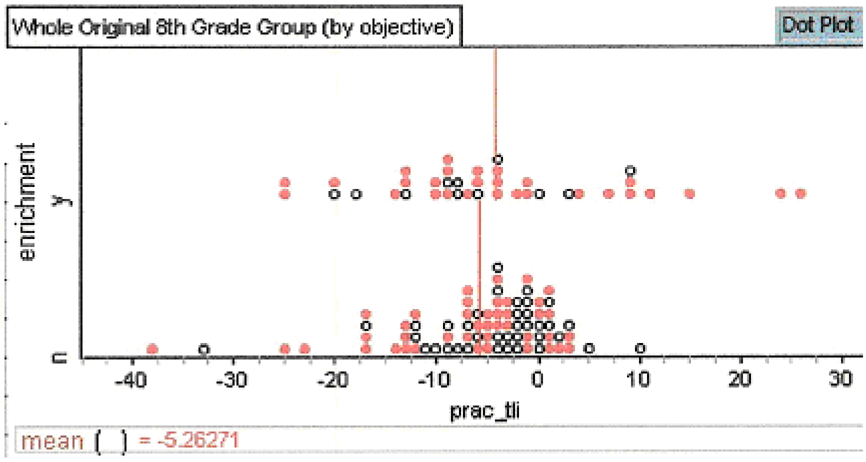


Figure 1. Graph shown to subjects during the interview task.

The preservice teachers were walked through particular elements of the graphs both in the pre-interview (January) and post-interview (May). It was explained that the data on the horizontal axis represented the improvement of each student from the seventh grade state exam to their eighth grade practice test. The mean improvement of each group, marked by a vertical line segment in each distribution, was pointed out. The overall mean (displayed as -5.26271 at the bottom of the graph) was interpreted to respondents as the average improvement of the entire eighth grade (that is, both groups combined) and it was also noted that the overall mean improvement lay in the negative region, meaning that generally the students had performed less well on their eighth grade practice test than they did on the seventh grade state exam.

In our choice of task we considered it important to make the context and task as authentic as possible in order to examine the prospective teachers' responses in a situation close to what they would encounter in their professional life. Therefore, rather than use hypothetical data constructed to emphasize a particular aspect of the distributions, actual data from a local school was used. These somewhat "messy" data made the task, and hence our analysis, more difficult. Yet, because authentic school data is rarely "clean" this setup provided the benefit of examining how the prospective teachers would interpret actual (and messy) school data. We recognize that unintended elements of this particular representation may have influenced the subjects' thinking about the task (Kosslyn, 1994). We had the teachers consider the "improvement" of students since improvement is a natural construct in teaching, and means were marked because this is a common method for comparing distributions, giving the subjects a potential starting point for their discussion. In addition, it allowed for insight into whether the prospective teachers would interpret a small difference in means deterministically, or if they would expect some variability between the means (Makar & Confrey, 2004).

4.5. ANALYSIS

The videotaped interviews from all twenty-two subjects were transcribed and then the content analyzed and coded to find the categories of concepts that emerged from the data. General categories were initially sought through open coding to isolate concepts that might highlight thinking about variation and distribution, and those passages identified by these codes underwent finer coding resulting in eighteen preliminary categories. Since codes were not predetermined, but rather allowed to emerge from the data, this portion of the analysis was not linear and underwent several iterations of coding, requiring a back-and-forth analysis as codes were added, deleted or combined. Commonalities and differences were examined in passages coded under each node to better describe and isolate the category, determine dimensions and distinctions among participants' descriptions, and locate exemplars. Although the data from all twenty-two original subjects were coded to determine categories, dimensions, and exemplars, only the seventeen subjects with complete data sets (both pre- and post-interviews) were used in quantitative descriptions.

5. RESULTS

In this section we first overview key categories of standard terms and concepts that the respondents, prospective teachers, used when comparing distributions. Next, categories of nonstandard language and terms are described, referring to two separate but overlapping areas: variation (e.g., *spread*) and distribution (e.g., *low-middle-high*, *modal clumps*). We later refer to these two types of non-standard categories as "variation-talk". Finally, relationships found within and between the categories of standard and nonstandard language categories will be summarized. As explained, the primary goal of the study is to provide a rich description of prospective teachers' language when discussing variation; hence, the information from the January and May interviews is usually combined. Changes in respondents' articulations from the first interview to the interview conducted after the course are noted as well but should be interpreted with caution in light of the small number of respondents and other factors noted later.

5.1. STANDARD STATISTICAL LANGUAGE

This subsection describes the conventional statistical language used by the respondents. Table 1 summarizes the percentage of respondents articulating each category of standard statistical terms and descriptions in their responses. As can be seen in the table, multiple types of standard expressions were used by respondents (i.e., percentages sum to more than 100%) and overall, nearly every subject included at least one type of standard statistical description in their response (94% in January and all respondents in May). Most respondents included the proportion (or number) improved or the mean in their descriptions and the inclusion of standard statistical terms in their responses increased in all categories by the end of the course.

Table 1. Percentage of respondents using standard statistical language (N=17)

Category	January	May
Proportion or number improved	59%	65%
Mean	53%	88%
Maximum/Minimum	29%	41%
Sample size	18%	47%
Outliers	18%	41%
Range	12%	47%
Shape (e.g., skewed, bell-shaped)	12%	35%
Standard deviation	0%	12%
Overall	94%	100%

Proportion or number improved

The most common comparison the respondents made in January was through reporting on the students in each group whose scores improved or dropped, with three respondents describing improvement as the sole element in their comparison. The prevalence of discussing improvement is not surprising given that the data in the task measured the improvement of students' scores.

In most cases, respondents split the groups into two—improved or not improved—as exemplified by Hope (all names are pseudonyms):

Hope: Some people have improved. A lot of people. (Jan)

In May, mention of the proportion of students improving persisted in the respondents' comparisons of distributions. The respondents were more likely to quantify their descriptions and none of them relied on proportions as their sole piece of statistical evidence.

Charmagne: There seem to be, like a split between, um, those who improved and those dropped, like, sort of, off the 50-50 split. (May)

One could argue that a focus on the proportion of students who improved does not necessarily imply that respondents are visualizing the variation in the data, nor seeing the distribution as an aggregate.

Means

Only about half of the respondents interviewed used the average in their descriptions in January, despite the fact that the means for each group were marked on the figure and also pointed out by the interviewer when describing the task. Use of the mean in comparing the distributions ranged from a brief mention to a major focus of their discussion. For example, two of those who mentioned the mean did not use any other statistical descriptions to compare the distributions:

Mark: Well, it looks to me like, uh, the group that did the Enrichment program overall has a better, uh, improvement even though it's not really even- [an improvement].

I: Okay. … And what are you basing that on?

Mark: Uh, cause you. I think you said that this line was the mean? … So, uh, I was looking at that. (Jan)

José: It seems about even. I mean, they didn't decrease by that much, compared to the other [group]. … I don't even know what that would be, a point between their mean and their mean? (Jan)

In previous work (Makar & Confrey, 2004), the authors noticed that while some respondents had a deterministic view of measures, others indicated some tolerance for variability in means, as did two respondents in this study who recognized the effect a small sample could have on the variability of the mean. The first excerpt below comes from Angela, a teacher with no formal training in statistics, whereas the second teacher, Janet, was a post-graduate student with a strong background in statistics.

Angela: Um, well, it's, I guess, obvious, I guess that. As this group [Enrichment], they did improve more, just I mean, because their average is better. But it's not a huge dramatic difference. … I mean, there's not as many in the Enrichment program [as the non-Enrichment] and they did improve more, but yet, I mean, I mean out of a smaller group of number. So their mean, I mean, comes from a smaller group. … I mean, if there were more kids, their average might have been different. (Jan)

Janet: So the Enrichment class did have a higher mean improvement, higher average improvement, uh, but they had a smaller class. Um, I don't know what else you want me to tell you about it.

I: You said they had a smaller class, is that going to have any-

Janet: A smaller sample size can throw things off.

I: How's that?

Janet: (laughs) Um. The, with a, a larger population the outliers have less of an effect on the, on the means than in a smaller sample. So it doesn't, um, I don't remember how to say it, it doesn't, uh, even things out as much. (Jan)

Janet's initial statement "I don't know what else you want me to tell you about it" may imply that she saw a difference in the means, but little else worth discussing. Kathleen, who had also used some statistics before in science, recalled comparing means there:

Kathleen: The mean [Enrichment] was a little bit higher than the, the group who didn't, who didn't take the Enrichment class. And I don't know if that would be statistically higher, but-

I: What do you mean, 'statistically higher'?

Kathleen: Like if you, if you ran statistics on it. Like a t-test or something.

When pressed further, Kathleen went on to explain in more detail:

Kathleen: If you, um, if you normalize the data, and um, brought them in together. In fact, once you normalize it for the number of students in this case [Enrichment group] versus the number of students in this case [non-Enrichment group] and brought them, like, closer together for the, for the number of students, and normalized it, then I think the difference [in the means] wouldn't be as great. (Jan)

Although through further probing she was unable to articulate what she meant by "normalize the data", it seems likely that Kathleen was referring to the dependence of sample size on key outcomes of the Central Limit Theorem to compare means with sampling distributions. In the May interviews at the end of the course, nearly all of the respondents mentioned the means, often with more specificity:

Anne: Well, it looks like the students in the Enrichment class, on average, um, improved, or didn't decline as much as the ones in the regular class. Um.

I: And what are you basing that on?

Anne: The means. Uh, the regular class is down by negative, uh, seven, six, minus six. And the Enrichment on average is at minus, um, is that three? (May)

While mean and percentage improvement are important considerations when determining the effectiveness of a class targeted to help students improve their test scores, our hope was that the respondents would do more than just reduce the data and compare means or percentage improvement as their sole method in determining how well the Enrichment program may have worked. Instead, we sought a more robust understanding of the context and an examination of the whole distribution in describing their comparisons.

Outliers and Extreme Values

Another common notion in comparing the relative improvement of each group described by the respondents at the beginning of the course arose through examination of outliers and extreme values. For example,

Andre: Well, it seems like with a few outliers here and a few outliers here, they're pretty similar, um, in terms of how much they changed. (Jan)

Andre had previously studied statistics. The descriptions by other students in the course with no statistics background were less precise:

Gabriela: There's only, like these two out here that have actually, like, greatly intensely improved. (Jan)

Gabriela focused on not just the criterion of *whether* students improved, but qualified it with *by how much*, suggesting that she was seeing the upper values of the distribution and not just whether or not students improved. Note that Andre and Gabriela are not focusing on individual points, but on a set of values at the high or low portion of the graph (e.g., those who "greatly intensely improved").

Shape

The interview task likely did not illicit a need to formally describe the shape of the distributions (e.g., skewed, normal), so their summary here is brief. Traditional shape descriptions were unlikely in January and only somewhat more common in May:

Christine: The non-Enrichment group seems to be skewed to the left. Uh, which means that any outliers that they do have are in the way negative region. Um. The Enrichment group seems to be more normal. It's slightly skewed to the right, but not quite. (May)

Standard Deviation

No one in January and only two of the subjects in May made any mention of standard deviation, a traditional measure of variation, despite the fact that it was discussed in class and included in a homework assignment. In one of these cases, a teacher mentioned standard deviation, but not for any particular argument except to state its relative size in each distribution:

José: Probably the standard deviation is going to be, like, really large on this [Enrichment], compared to that [non-Enrichment], because this is pretty spread out pretty far. (May)

Another prospective teacher indicated that she knew the term, but implied that she did not see it as useful in comparing the relative improvement of the students in the Enrichment program with those who were not, stating a few minutes into her interview:

Charmagne: Um. Yeah. There is more variation in the Enrichment class. This seems to be kind of mound-shaped also. So. I mean. Probably like 65% is in one standard deviation, [laughs] I'm just babbling now. Did I answer the question yet? (May)

From these, the only two examples of the subjects mentioning standard deviation, it would appear that the notion of standard deviation as a measure of variation did not hold much meaning for the respondents. Both of these excerpts pair the use of standard deviation with other less standard expressions that described the variation ("pretty spread out") or shape ("more variation" vs. "kind of mound-shaped"). This may imply that these less conventional descriptions of variation aided them in making meaning of standard deviation.

Range

It was often difficult to tease out notions of variability from descriptions of measures of variation, particularly when the respondents used terms like "range". Ten respondents used the word "range" during the interviews (2 in January and 8 in May, with no respondents in common). In almost every case, their use of it was linked to notions of either measure or location (an interval).

Carmen: The Enrichment class definitely had a better performance since most people are concentrated in this area, whereas you have a wider range and even a very good amount of points that they improved on. … I think it is working, yeah. Because you have a just wider range, whereas everyone was kind of close in on their improvement here. Uh. With the wider range, um, I would say it's working at least for some of the students. Because in the non-Enrichment, no one seemed to improve that much. … Because there's not a range here [the upper portion of the scale]. (Jan)

Carmen's use of the word "range" four times in this passage communicates range as an interval of values in the distribution rather than as a measure. First, she contrasted "wider range" with "concentrated in this area" and "close in … here" giving the impression that she was expressing that the data were spanning a greater part of the scale at a particular location. Next, her suggestion that the wider range implied it was "working at least for some of the students" indicates that it was located in the upper part of the scale, unlike the data for the non-Enrichment group where she said "there's not a range here [the upper portion of the scale]." Brian also used the term "range" to indicate the scale:

Brian: It seems pretty evenly distributed across the whole scoring range. (Jan)

In the May interviews, the use of "range" was more common than in January, even though it wasn't a term we made use of formally in the class. In almost every case, the term "range" either meant an interval, as in the case of Carmen and Brian above, or a measure, like April:

April: The distribution, um, like the lowest the scores in the distri-, the length of the distribution, see this one starts, it's. [pause] … This one is about negative, almost negative forty, I'd say. And this one goes up to ten. So, that's about 50. And this one's about negative 25 and this one's right about 25, a little more, so that's about 50. So, I guess the range is about the same. (May)

Gabriela: There's a lot less of them improving in the Enrichment program, but it's still better that they go off by about five or ten points … then for them to have gone off by forty or twenty. Still kind of in this range. (May)

April's use of the term range is more numeric whereas Gabriela's use appears to indicate a segment of the scale. One difficulty may be that in school mathematics, the term "range" is usually related to a function and defined as a set, almost always an interval on the real number line. In statistics, however, the term "range" is a measure—the absolute difference between the minimum value of a distribution and the maximum value. By using the same term to indicate a set and a measure, we begin to see where the distinction between objects and measures become murky in statistics. In school, the distinction between a geometric object (like a polygon), a measure of it (its area or perimeter), and a non-numerical attribute or categorization of it (closed or convex) is made clear. In teaching statistics, we have not emphasized a clear distinction between an object (e.g. a distribution), a measure of it (its mean or interquartile range), and an attribute (e.g. its shape). This may cause some problems for students trying to make sense of statistical concepts.

5.2. NONSTANDARD STATISTICAL LANGUAGE

This subsection documents the phrases used by the respondents to articulate statistical concepts which could not be categorized as standard statistical terminology. Two dimensions of nonstandard statistical language emerged from the observations made by the respondents (Table 2): *spread* and *distribution chunks*. Nonstandard statistical phrases were categorized into one of these two categories (i.e., they were mutually exclusive), however the dimensions of spread and distribution, as we will show in subsection 5.3, were related. Similar to the results of the use of standard statistical language in the previous subsection, many respondents included both types of nonstandard statistical descriptions (i.e. percentages sum to more than 100%). As can be seen in Table 2, most respondents (53%) included some mention of distribution chunks in their interviews in January and this percentage increased somewhat in May. Although few respondents (35%) discussed the spread of the data in the January interviews, this percentage increased markedly in May. Overall, in both the January and May interviews, the majority of respondents included some kind of nonstandard statistical language in their responses and this percentage increased from the beginning to the end of the course.

Table 2. Percentage of subjects making observations in two dimensions of nonstandard language (N=17)

Dimensions	January	May
Spread	35%	59%
Distribution chunks	53%	65%
Overall	59%	76%

***Expressions of variation:* Clustered *and* spread out**

Here we will document the nonstandard statistical language used by respondents in the interviews that capture their articulation of variation. Their words encompassed a diverse range of language, but the concepts they articulated were fairly similar.

Some respondents used the word *clustered* to describe the relative improvement of each group, like Andre and Margaret, both older college students with previous statistical experience:

Andre: I don't know what to make of this, actually, because as far as, like, it seems to me to support little difference between the Enrichment group and the other group. Because. Um. Both groups are kind of *clustered* around the same area. (Jan)

Margaret: [The non-Enrichment data] are more *clustered*. So where there's little improvement, at least it's consistent. This [Enrichment] doesn't feel consistent. First impression.

I: And you're basing this on?

Margaret: The *clustering* versus, it's like some students reacted really well to this, and some didn't. But it's more *spread out* than this grouping. (Jan)

Andre's use of the term "clustered" highlights his observation that the location of the modal clump in the two distributions overlapped. On the other hand, Margaret's initial description of "clustered" is paired with a notion of consistency, a concept closely related to variability (Cobb, 1999). She goes on to include it in contrast to being "more spread out", a phrase commonly associated with variation (Canada, 2004; Bakker, 2004; Reading, 2004). Five other respondents made use of the phrase "spread out" during the interviews. A few excerpts are given below.

Brian: It seems to be pretty *evenly distributed* across the whole scoring range. Like from about 30 to [negative] 25, it appears pretty *evenly spread out*. (Jan)

Janet: They seem pretty *evenly distributed*, … fairly *evenly spread out*. (Jan)

April: This distribution is more *skewed to the left* and this one is more *evenly spread out* … more of an *even distribution*. (May)

The respondents who used the phrase "spread out" also accompanied it with the word *evenly*. They may be expressing that the data were dispersed fairly equally throughout the scale of the distribution, particularly given its common pairing with the phrase *evenly distributed* or *even distribution* in all three cases above. This context gives *spread out* a meaning related to the *shape* of the distribution, particularly given the contrast April made between *skewed left* and *evenly spread out*. Carmen's description below gives the phrase a similar meaning:

Carmen: It's more *spread out*, the distribution in the Enrichment program, and they're really kind of *clumped*, um, in the non-Enrichment program. (May)

Given this interpretation, we turn back to and expand Margaret's excerpt to re-examine her use of the phrase *spread out* contrasted with *clustered*, which is similar to Carmen's *clumped* above:

Margaret: It's interesting that this is not, that this is not, um, this is much more *spread out* than this group, so. I mean, first impressions. … These are more *clustered*. So where there's little to no improvement, at least it's consistent. This doesn't feel consistent. First impression.

I: And you're basing that on?

Margaret: The *clustering* versus, it's like some students reacted really well to this, and some didn't. But it's more *spread out* than this grouping. Is, am I saying that okay? (Jan)

The term *spread out* in all these cases appeared to point to an attribute of the distribution akin to shape; we would argue that *clustered* and *clumped*, phrases that accompanied *spread out*, could also be attributes that describe the shape of the distribution, similar to Konold and his colleagues' (2002) notion of *modal clump*.

How is *the spread* related to *spread out*? Given their similarity, how does its use as a noun compare with its use as an adjective?

Carmen: If you were just to, you know, break the distribution in half. Kind of based on the, the scale, or *the spread* of it, it just seems that, you know, the same amount of students did not improve in both. (May)

Here, Carmen indicates that she is using the noun *the spread* as an indication of length by her pairing of *the spread* with *the scale*. Rachel, in her interview in May, first compared the means of the Enrichment and non-Enrichment groups, then turned to the range, and finally, below, finishes the interview by discussing the way the distribution looked:

Rachel: It's more *clumped*, down there in the non-Enrichment. And kind of more *evenly distributed*. [Points to Enrichment] … Let's see. Range and *spread*. That's what I always first look at. And then average. (May)

Margaret: [The Enrichment distribution] has a much wider *spread* or *distribution* than this group [non-Enrichment]. (May)

In all three of these cases—the only ones where *spread* is used as a noun—Carmen, Rachel, and Margaret convey a meaning of spread as a visual (rather than numerical) attribute of a distribution. Rachel uses the word *spread* to categorize her description of two contrasting terms *more clumped* and *more evenly distributed*. Note that she also distinguishes her notion of *spread* as different than *range* (as a measure), which she discussed earlier in the interview. Margaret directly pairs the word *spread* with its apparent synonym *distribution*, although she likely was using the word distribution in a more colloquial sense.

Another phrase that conveyed similar meaning to *spread* was *scattered*, as used by three respondents:

I:	The first thing I want you to do is just to look at those two and compare the two in terms of their relative improvement or non-improvement. We're trying to determine if the program is working.
Hope:	Well, it's doing something.
I:	What do you mean?
Hope:	I mean, they're more *scattered* across, these guys. … It's helping a little.
I:	Okay. And you're basing that on?
Hope:	On. Well, there's more *grouped* right here. … But you have guys spanning all the way out to here, so it's helping. … It's helping, it's *scattering* them more, it seems. Instead of them all having, so *grouped* together. (Jan)

Hope's descriptions are akin to those we heard above with phrases like *spread out* and *clustered* and *clumped*. Janet's and Anne's uses were similar:

Janet: There's Economically Disadvantaged kids pretty much *scattered* throughout both graphs. (May)

Anne: I mean these are all kind of *scattered* out almost evenly. Whereas these are more *bunched* up together. (May)

If we substitute *scattered* with *spread out* above, notice how the meaning doesn't appear to change. Also note Anne's contrast of *scattered out* with *bunched up*. One more pairing may also be included here:

June: It seems that the people that weren't in the Enrichment seems to be all *gathered* from the zero and the negative side compared to the people that were in the Enrichment program because this is kind of *dispersed* off and this is like, *gathered* in the center. (Jan)

From these excerpts, we can collect a set of terms under the umbrella *spread* that indicate similar notions: *spread out*, *scattered*, *evenly distributed*, *dispersed*. Antonyms include *clumped*, *grouped*, *bunched up*, *clustered*, *gathered*. Three other terms, *concentrated*, *tight*, and *close in*, appeared too infrequently to compare, but conjectured to be similar. Note that many of the respondents' articulation of variation accompanied or integrated distribution shape, argued by Bakker (2004) to be a key entry point for understanding variation. The next section will look at language the respondents used to describe distribution – what we call *distribution chunks*, or context-relevant distribution subsets.

Expressions of distribution: Meaningful chunks

This subsection will tackle the second dimension of nonstandard statistical language used by the prospective teachers in the study: observations regarding the distribution. Although our goal in the study was to investigate the way the prospective teachers articulated notions of *variation*, we found that many of their observations relating to the distribution were also rich with notions of variation. In this subsection, we will discuss three particular kinds of observations of the distribution emerged from the data: triads, modal clumps, and distribution chunks.

Triads. Although describing the data by the number or percentage who improved was very common, not all of the respondents used this criterion to split the distribution into two categories: improvement and non-improvement. Other respondents partitioned each distribution into triads—improving, not improving, and "about the same"—as Maria and Chloe did in January:

Maria: Well the people who weren't in the Enrichment program they did score lower, and the one in the Enrichment program, their scores are kind of varied—some of them did improve, others stayed about the same, and some decreased. … The ones who didn't take [Enrichment] basically stayed the same. There was no real improvement. There's maybe, maybe a few that did, but not so many. While in the Enrichment, there was a lot that did improve. (Jan)

Chloe: Well, uh, it seems like the people that aren't in the Enrichment program, they're staying the same or even getting worse off at, with the next test. But the people in the Enrichment program it looks like, it looks like it's evened out. Like you have some doing better, some doing worse, and some on the same level. There's a little bit more doing worse, but you still got those few that are still doing better. (Jan)

Here, Maria and Chloe partitioned the distribution into three categories: those who improved, those who "stayed about the same" and those whose scores decreased. This seems to indicate that they saw more than just the students who scored above and below an absolute cut point of zero (as described in section 5.1) and that the demarcation between improving and not was more blurred. Their decision to split the distribution into triads ("some doing better, some doing worse, and some on the same level") is consistent with Bakker's (2004) finding that students often naturally divide the distribution into three pieces – low, middle, and high. Viewing the distribution in three pieces is one way that the respondents may have simplified the complexity of the data. This supports work by Kosslyn (1994) who argues that the mind is able to hold no more than about four perceptual units in mind at one time.

Modal clumps. Beyond recognition of the distribution into three chunks–low, middle, and high–several of the respondents focused specifically on the middle portion of the distribution, seeing what Konold et al. (2002) refer to as the *modal clump*:

> To summarize their data, students tended to use a 'modal clump,' a range of data in the heart of a distribution of values. These clumps appear to allow students to express simultaneously what is average and how variable the data are. (p. 1).

A diverse set of expressions indicating awareness of a *modal clump* emerged from several of the respondents' responses:

Chloe: It seems like the most, *the bulk* of them are right at zero. (Jan)

Janet: [The Economically Disadvantaged students] seem pretty evenly distributed. I mean, from this bottom group here, they, the two kids with the highest scores are not Economically Disadvantaged, but that's just a, they don't have that much improvement above the others, above *the main group*. (Jan)

Hope: It's kind of flip-flopped *the big chunk* of this one is on this side, *the big chunk* of this one is on this side. (Jan)

Gabriela: *The majority* of the ones that took Enrichment anyways are still more in the middle. [pause] Or they stayed the same, or they got worse, so I would say that just it's not an effective program. (Jan)

The use of terms like *the bulk of them*, *the main group*, *the big chunk*, and *the majority*, by these respondents all indicate an awareness of a modal clump. In addition, Gabriela's observation that the distributions overlapped factored into her decision about whether she thought the Enrichment program was effective. This suggests that she may have been expecting a more distinct division between the two groups if, in fact, the Enrichment program was working. Gabriela seemed to be basing much of her tendency to split the distribution into three groups by focusing on the *scale* (position relative to zero) rather than on the notion of Bakker's (2004) *low-middle-high*. This perspective changed in her May interview as she added an additional caveat to her description:

> *Gabriela*: If the, if the mean will tell you, that give or, give or take five points [of a drop], that that's 'normal' or 'usual' for them, that then that's not cause to think that 'well, they're not improving' or the program's not working – if they stay around those [negative] five points. (May)

Gabriela's decision about whether the Enrichment program was working changed from her January to May interview based on her interpretation of the middle clump. In January, she indicated that the program wasn't working *because* the middle clump overlapped with the non-Enrichment group, and its location being below zero meant most of the students hadn't improved. In May, however, she argued that because a five-point drop was *typical* for the eighth graders as a whole, you could not use the fact that it was negative to argue against the success of the Enrichment program. Notice that she is using the middle clump here as a description of the average of the group. The notion of this clump persisted for others as well in May and its frequency increased, as evidenced by April:

> *April*: *The majority* of their sample size is on the right side of the mean. Um. And I'd say this is about even, maybe a little more on the opposi-, on the left side of the mean. (May)

Mention of proportion or number improved was very common in the respondents' reports in comparing the improvement of each group (more common than reports of average), and many of the respondents who compared the two groups based on proportion of improvement also saw a *majority* or *main group* clump; this may indicate that modal clumps are strong primitive notions of variation and distribution. The fact that almost all of the respondents used some sort of criterion to separate each distribution (into either two or three groups) also implies that modal clumps may be a useful starting point to motivate a more distributional view.

Distribution chunks. Not all of the respondents who described *distribution chunks* did so with the modal clump or as a way to split the group into low-middle-high. Several of the respondents examined a different meaningful subset of the distribution, most typically a handful of students who improved the most:

> *Anne*: It looks like a few students responded really well to the Enrichment class and improved their scores a lot. … [And] for this, these, this student in particular, but all of these [pointing to a group of high improvers], the Enrichment program worked. I would say that. (Jan)

> *Carmen*: It just seems like the majority of them didn't improve very much. … You still have these way up here—not just the fact that they, that these two improved so much [two highest in Enrichment]—but you have several that went way beyond the average, you know, they went beyond the majority that, of the improvement here. (Jan)

Carmen's descriptions in this early interview indicate she was not only paying attention to whether or not the improvement was above zero, but also made note of the outliers "these two [that] improved so much" and those that "went way beyond the average". In examining *the majority*, and two different groups of high values, she indicates that she is changing her field of view from a fixed partitioning to a more dynamic or fluid perspective. At her May interview, she continued this facility with moving between distribution subsets, but was more specific in her description. A few of the subsets she mentions in this longer exchange are noted (Figure 2):

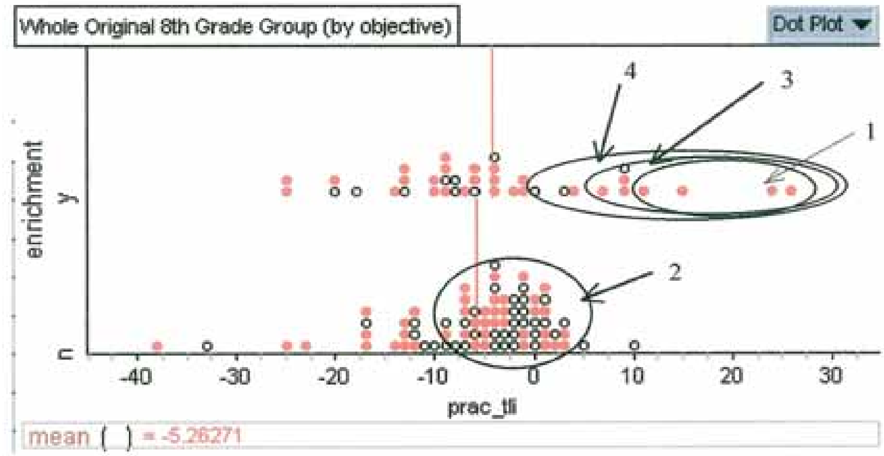


Figure 2. Carmen's description with several meaningful subsets of the distribution circled.

I: Compare the improvement of the students who were in the Enrichment program with the students that weren't.

Carmen: Um, well, … the student with the highest improvement in the Enrichment program was about 16 points above the student with the highest improvement in the regular program. Um. And then, uh, this clump [upper Enrichment], there are others that are higher than the highest improvement, too. I mean there's about four that, um, improved more than the student in the regular program with the highest score [Figure 2, circled portion 1]. Um. It looks like, well there are more people, more students in the non-Enrichment program, and, um, and the non-Enrichment program, um, they also scored considerably lower. I guess the improvement was considerably lower than the students in the Enrichment program. So. I don't know. …

I: Okay. Um. So in your opinion, would you say the program is working? Should they continue it?

Carmen: Um. Well. It seems that it, it is working, that, um. I mean, all of these students improved a lot more than, uh, this big clump of students here [circled portion 2] in the non-Enrichment program, um, but on the same token, it looks like there was about the same that didn't improve in both programs.

I: Do you mean the same number or the same percentage?

Carmen: The same number. Around. Um. … I think, I think it is working and they should probably continue it because of the ones that improved, you know, well there's four that improved much greater than the, than the one that improved ten, by ten in the non-Enrichment group, um, but then there were, there were about eight that improved, uh, more than majority of the non-Enrichment group [circled portion 3]. So even though it wasn't, well, probably one-third, um, showed a considerable improvement [circled portion 4] and I, I would think that that's worth it. (May)

Here, Carmen's ability to interpret and make meaning of the distributions goes beyond recognition of the distribution as the frequency of values of a variable (improvement scores). Besides a quantification of her view of outliers (16 points above the maximum of the non-Enrichment group), one could argue that Carmen is seeing the set of high values in the distribution as more than just individual points, but as a contiguous subset of the distribution. As demonstrated above by the circled areas on the distribution, Carmen was also demonstrating her ability to see several *distribution chunks*, with dynamic borders. This perspective of chunks (as a subset rather than individual points) is more distribution-oriented and indicates that examining chunks, beyond just the minimum and maximum, may be a useful way to encourage teachers to adopt a more distribution-oriented view of the data. Tentatively, the notion of distribution chunks seems to fit somewhere between a focus on individual points and a view of the distribution as a single entity or aggregate (see Konold, Higgins, Russell, & Khalil, under review). As well as having a perspective of distribution as both whole (aggregate) and part (subset), Carmen was also clear in her ability to articulate meanings of both whole and part in terms of the context. For example, she recognized the cut point for "improvement" as values above zero and used the clumps in the distribution to identify the majority of students' improvement scores.

5.3. RELATIONSHIPS BETWEEN CATEGORIES

In this subsection, we briefly examine two perspectives of relationships between and among standard and nonstandard language use, and provide data about the use of language by respondents in January and May. First, we compared the number of standard and nonstandard terms given by each respondent. Recall that the percentage of respondents using standard or nonstandard language changed little during the course (Table 1 and Table 2). However, the mean *number* of distinct terms used by the respondents showed a slightly different pattern from January to May, as shown in Table 3. There was a significant difference between the mean number of standard terms used in May and in January ($t_{16} = 2.84$, $p = 0.01$) and Table 3 shows that the mean number of different standard statistical terms used by respondents increased between January and May. One explanation for this is that the respondents learned (or reviewed) conventional statistical terms during the course. There was no significant difference between January and May in the case of non-standard terms ($t_{16} = 0.77$, $p = 0.46$). In addition, looking at the total numbers of standard and non-standard terms used by respondents in January and May combined, the difference between the mean numbers of terms used was not significant ($t_{16} = 1.73$, $p = 0.10$). That is, in general, respondents likely used no more standard statistical terms than nonstandard statistical observations. Note, however, that there is greater within-group variation for nonstandard terms, suggesting nonstandard statistical language was used less consistently across subjects than standard statistical language.

Table 3. Mean (standard deviation) number of standard and nonstandard terms used by respondents (N=17)

	Standard Observations	Nonstandard Observations
January	2.12 (1.76)	1.94 (2.19)
May	3.88 (1.36)	2.47 (2.07)
Change (from January to May)	1.76 (2.56)	0.53 (2.85)
Total (January and May)	6.00 (1.84)	4.41 (3.16)

Next, we looked at nonstandard language use and whether the dimensions of nonstandard statistical language (variation and distribution) are overlapping or independent. The two dimensions of nonstandard language were suspected to be related. For example, describing a distribution's variation using the term "clumped" also described the distribution as mound-shaped. Because the nonstandard descriptions of variation often contained distribution characteristics, we wondered to what extent articulation of these concepts overlapped. Most of the subjects did not describe the distributions in the same way across interviews. For example, if they mentioned notions of spread in January, it was no more or less likely that they would mention spread in May. We thought it safe, therefore, to combine the types of responses from January and May in order to investigate, as a whole, whether notions of spread and distribution could be considered independent or correlated. Table 4 displays the number of respondents articulating nonstandard expressions of variation and distribution in their interviews. Much of the data in the table falls along the diagonal which implies that most respondents tended to describe either both the variation and distribution in the data or neither. A Fisher's Exact Test (Cramer, 1997) was used to investigate the relationship between these two dimensions and was found to suggest an association between nonstandard expressions of variation and distribution ($p = 0.02$). An interpretation of this could be that the respondents tended to "see" variation and distribution in the data graphs (or not) together. Another possibility is that nonstandard statistical language naturally integrates these concepts.

Table 4. Number of respondents incorporating nonstandard expressions of variation and distribution (N=34)

	Nonstandard Expressions of Variation		
Nonstandard Expressions of Distribution	Yes	No	TOTAL
Yes	13	7	20
No	3	11	14
TOTAL	16	18	34

6. DISCUSSION

The goal of the study was to gain insight into the ways in which the respondents, prospective teachers, expressed notions of variation in comparing data distributions in a relevant context. The task given to the teachers asked them to compare the relative improvement of test scores between two groups of students. Two categories of statistical language emerged from the teachers' descriptions: standard statistical language and nonstandard statistical language. The diversity and richness of their descriptions of variation and distribution demonstrated that the prospective teachers found many ways to discuss these concepts, and that through their nonstandard language, they were able to articulate keen awareness of variation in the data. Two dimensions of nonstandard language were found—observations of spread (variation) and observations of meaningful chunks (distribution). These dimensions overlapped, indicating that either the respondents saw these two notions (or not) together, or that the nonstandard language naturally integrated notions of variation and distribution. In addition, no overall quantitative differences were found between the prospective teachers' use of standard and nonstandard statistical language.

This section will discuss two outcomes of the study: characteristics of nonstandard statistical language or "variation-talk", and elements of the structure of standard and nonstandard statistical language. We will close with a reflection on limitations of the study.

6.1. CHARACTERISTICS OF "VARIATION-TALK"

The subjects in the study expressed concepts of variation using nonstandard language in a variety of ways; we call these ways of articulating variation *variation-talk.* The variation-talk used by these prospective mathematics and science teachers was not so different than the language that emerged in other recent studies of learners' concepts of variation and distribution (Bakker, 2004; Canada, 2004; Hammerman & Rubin, 2004; Reading, 2004), however these studies did not look at nonstandard language systematically and employed other comparison tasks.

The results in this study classified the prospective teachers' *variation-talk* into four types: *spread*, *low-middle-high*, *modal clump*, and *distribution chunks*. The nonstandard language used by the teachers to express *spread*—clustered, clumped, grouped, bunched, gathered, spread out, evenly distributed, scattered, dispersed—all past participles, highlighted their attention to more spatial aspects of the distribution. These terms took on a meaning that implied attention to variation as a characteristic of shape rather than as a measure. This is consistent with Bakker's (2004) description of shape as a *pattern of variability*. In contrast, concepts of variation in conventional statistical language are articulated by terms like *range* or *standard deviation*, both of which are measures.

The other three types of "variation-talk", all nouns, focused on aspects of the variability of the data which required the prospective teachers to partition the distribution and examine a subset, or *chunk* of the distribution. Similar to Hammerman and Rubin's (2004) teachers, the prospective teachers in this study simplified the complexity of the data's variability by partitioning them into bins and comparing slices of data, in this case three slices: *low-middle-high*. Although simple, this is a more variation-oriented perspective than responses that took into account only the proportion or number of students who improved on the test. This approach is consistent with the findings of Bakker (2004), who argued that partitioning distributions into triads may be more intuitive than the more conventional partitioning into four, as in the box plot. The awareness of learners' tendency to simplify the complexity of the data into low-middle-high bins motivated the creators of Tinkerplots (Konold & Miller, 2004) to include the "hat plot" (Konold, 2002) in the software, a representation where users can partition the data into thirds based on the range, the average or standard deviation, the percentage of points, or a visual perspective of the data.

In some excerpts, teachers argued that the overlap of the middle slices was evidence that the Enrichment class was not effective. The focus on the *modal clump* (Konold et al., 2002) in data has been a consistent finding across several studies of both students and teachers, reinforcing both the intuitive nature of seeing variability through slices-in-thirds, and recognizing the potential of using the notion of a modal clump to encourage learners to move from a focus of individual points towards a focus on the *aggregate* of a distribution. Having a lens of a distribution as an aggregate, as opposed to a set of individual points, allows for concepts such as center to be thought of as a characteristic of the distribution rather than as a calculation derived from individual points (Bakker, 2004). In addition, locating a *modal clump* allows the learner to simultaneously express their visualization of the center and spread of the data (Konold et al., 2002), again highlighting its relational nature.

Finally, the use of *other meaningful chunks* by several of the prospective teachers demonstrated their ability to focus on the variability of the data by examining particular subsets of the distribution. This category contained responses as simple as comparing the outliers of one distribution to the "majority" of the other. A more complex visualization of distribution subsets was articulated by Carmen who used several different distribution subsets, with dynamic borders, as evidence for the effectiveness of the Enrichment program. Her facility to fluidly manipulate borders of these subsets highlighted her ability to visualize variation in the data. In addition, each of her *meaningful chunks* was tied back to the context of the problem, indicating that she was able to use them to make meaning of the situation. Although Konold and Bakker have encouraged the conceptualization of a distribution as an aggregate, the articulation of distribution chunks by these teachers suggests that they are thinking of distribution chunks as *mini-aggregates* of the data. Kosslyn (1994) argues that our minds and eyes work together to actively group input into perceptual units that ascribe meaning. We would argue that in articulating subsets of the distribution, the prospective teachers are communicating the perceptual units they are seeing in the data.

The three perspectives of seeing partial distributions—triads, modal clump, and distribution chunks—indicate that there are more than just the two perspectives of distribution that are usually discussed in the literature: single points and aggregate. This third perspective—partial distributions or "mini-aggregates"—deserves further research to investigate the strength of its link to statistical thinking about distributions. This study has highlighted that prospective teachers may use descriptions of partial distributions not only to articulate rich views of variation, but also to use these distribution chunks in meaningful ways that could not be captured using conventional statistical terminology. Even though we classified expressions of variation separately from notions of distribution, this separation was somewhat artificial in that all four types of "variation-talk" expressed a relationship between variation and distribution.

6.2. THE STRUCTURE OF STANDARD AND NONSTANDARD STATISTICAL LANGUAGE

The nonstandard language used by the prospective teachers by its very nature *integrates* the important statistical ideas of variation and distribution, capturing and implying cognitive relationships between notions of center, dispersion, and shape. In contrast, the standard statistical language used by the prospective teachers was by its very nature less relational, with a tendency to express important ideas in statistics as conceptually separate. Terms such as mean, standard deviation, and skewed distribution describe the center, dispersion, and shape of a distribution but they do so in isolation. The overuse of this standard terminology, at least early in learning statistical concepts, may encourage learners to maintain a perspective that statistical concepts are isolated "bits" of knowledge rather than information that can provide insight into relationships in the data. Several difficulties of learning statistics that are documented in research are consistent with a perspective of statistics as isolated facts. For example, seeing data as a set of isolated points rather than developing a "propensity perspective" (Konold, Pollatsek, Well, & Gagnon, 1997), lack of intuition in stochastics (Shaughnessy, 1992), or focusing instruction on calculations, isolated procedures, and graph characteristics in statistics instead of on drawing meaningful conclusions from data (Pfannkuch et al., 2004) may all be reinforced by the *overuse* (too much, too soon) of conventional statistical language.

The process of integrating rather than separating concepts in statistics has been shown to be a productive avenue for developing statistical thinking and reasoning (Konold et al., 2002; Konold & Pollatsek, 2002; Bakker, 2004). This does not imply that one should "teach" nonstandard statistical language as a means to encourage the development of students' intuition about statistics, but rather to encourage their sense-making by acknowledging and encouraging learners' own language. Had we acknowledged only conventional terms in our search for their articulation of variation, we would have lost many opportunities to gain insight into their thinking. For example, only two preservice teachers used the term "standard deviation" in their comparisons, and it can be argued that neither one used this concept to articulate meaning about variation.

An important lesson we learned in trying to categorize the preservice teachers' descriptions was how difficult it was to separate their observations of variation and distribution. For example, when April described the shape of one distribution ("skewed to the left"), she compared its shape with the *spread* of the other distribution ("evenly spread out"). Using the teachers' words, we could see that describing data as *clumped* or *spread out* said as much about the *distribution* of the data as it did about its *variation*. It was not surprising, therefore, when the responses in the two dimensions of nonstandard language (variation and distribution) were found to be correlated.

6.3. LIMITATIONS OF THE STUDY

Although we believe that the results of this study communicate a powerful message about the opportunities of listening to learners' nonstandard statistical language, several limitations of the study must be acknowledged.

- *Population*. The subjects in the study were prospective secondary mathematics and science teachers and the study is not directly generalizable beyond that population. While many of the results may seem to transfer to other groups (for example, practicing teachers), more research would need to be conducted in order to corroborate these results with other populations.
- *Particulars of the task*. The setting may have had strong influences on the responses that the prospective teachers gave during their interviews. Although the data were authentic, the task was not as it did not emerge from the teachers' own desire to know. Therefore, it is possible that the responses given by the prospective teachers may have mirrored an expectation of what they thought the researchers wanted them to say, particularly since the researchers were also their course instructors and the interviews were conducted as part of the course assessment.
- *Particulars of the graphic*. Elements of the graph could elicit responses that may differ with slight modifications. Recall, for example, that the means were marked on the figure and explicitly pointed out to the subjects at the beginning of the task. By drawing specific attention

to the means, we may have been tacitly communicating that the means should have been attended to in addition to or even instead of the prospective teachers' own way of comparing the two groups. The results of the Hammerman and Rubin (2004) study, for example, noted that the teachers in their study were not interested in using means to compare distributions but invented their own ways of making meaning of comparisons in the problems they were discussing.

- *Irregularities in questioning*. In a few of the interviews, the interviewer did not explicitly ask for a decision to be made regarding the effectiveness of the Enrichment program. This meant that comparisons could not be made systematically regarding the prospective teachers' use of evidence towards a decision, which could have elicited richer responses.
- *Generalizations about the course*. Given the unique nature of the course which the prospective teachers undertook, the study was not designed to make generalizations about the effectiveness of such a course in developing learners' statistical reasoning. Many of the compelling elements used in the course and the task were developed from local problems in the implementation of standardized testing. Instead, the study was designed to focus on variation-related language that preservice teachers use at different stages or levels of learning about statistics. Other reports of the larger research study communicate elements of the course that may have had an impact on the prospective teachers' thinking and learning (Confrey et al., 2004; Makar & Confrey, submitted; Makar, 2004).

7. CONCLUSION AND IMPLICATIONS

This study was built on pioneering work by researchers at TERC who first studied teachers' nonstandard statistical language (e.g., Russell & Mokros, 1990), and shows the breadth and depth of nonstandard language used by prospective teachers to describe variation. It follows that teachers need to learn to recognize and value "variation-talk" as a vehicle for students to express meaningful concepts of variation and distribution. This study has contributed to our understanding that preservice teachers majoring in math and science often articulate meaning in statistics through the use of less conventional terminology. Other research studies have shown that school children do so as well (e.g., Bakker, 2004; Reading, 2004; Konold et al., 2002). These studies have begun to articulate rich and productive learning trajectories to move students towards a more distribution-oriented view of data.

Although the preservice teachers in this study were using nonstandard statistical language, the concepts they are discussing are far from simplistic and need to be acknowledged as statistical concepts. Not recognizing nonstandard statistical language can have two pernicious effects. For one, we miss opportunities to gain insight students' statistical thinking. Noss and Hoyles (1996) explain the benefits of attending to students' articulation for creating mathematical meaning: "It is this *articulation* which offers some purchase on what the learner is thinking, and it is in the process of articulation that a learner can create mathematics and simultaneously reveal this act of creation to an observer" (p. 54, italics in original). Gaining insight into student thinking, therefore, is not the only benefit from attending to nonstandard statistical language. Noss and Hoyles also argue that through the process of articulation, students have opportunities to *create* meaning. Sense-making in statistics is an ultimate goal that is often neglected by more traditional learning environments. Valuing the diversity of students' "variation-talk" and listening to student voice (Confrey, 1998) may encourage teachers to shift from the typically procedure-focused statistics courses towards a focus on sense-making. This is an issue of equity if we are to acknowledge the diversity of students' ideas rather than just cover the content and label those who don't talk statistically as being *unable* to do so.

The other problematic effect of neglecting students' nonstandard statistical language is the tacit message that is communicated that statistics can only be understood by those who can use proper statistical talk. Lemke (1990) argues that the formal dialogue of science communicates a mystique of science as being much more complex and difficult than other subjects, requiring that we defer our own ideas to those of 'experts'. By doing so, Lemke argues

> The language of classroom science sets up a pervasive and false opposition between a world of objective, authoritative, impersonal, humorless scientific fact and the ordinary, personal world of human uncertainties, judgments, values, and interests … many of the unwritten rules about how we are supposed to talk science make it seem that way (p. 129).

Although Lemke here is talking about science, the situation in statistics is isomorphic. Countless students complain that statistics (which they call "sadistics") is not understandable to ordinary humans and lacks a connection to sense-making. By communicating statistics as accessible only to experts and geniuses, we are reinforcing a notion of statistics as a gatekeeper to powerful insights about the world, alienating students and denying opportunity to those who do lack the confidence in their ability to make sense of statistics.

Three major benefits come out of teachers' use of their own, nonstandard statistical terminology in describing, interpreting, and comparing distributions. First, they are using words that hold meaning for them and that convey their own conceptions of variation. In the constructivist perspective, knowledge is not conveyed through language but must be abstracted through experience. Nonstandard language carries a subjective flavor that reminds us of this. Through interaction with others, this subjectivity becomes *intersubjective*—one's meaning is not identical to another's, but through further explanation, our meanings become more compatible with the language of our peers (von Glasersfeld, 1991). Second, everyday uses of language are more accessible to a wide variety of students, allowing multiple points of entry to statistical concepts while encouraging teachers to be more sensitive to hearing rich conceptions of variation in students' voice (Confrey, 1998), words that may allow students easier access to class discussions. This is a more equitable, more inclusive stance; one that is contrary to the conception of mathematics (or statistics) as a gatekeeper. Third, if the goal is to provide students with experiences that will provide them with a more distribution-oriented view of data, then nonstandard statistical language that emerges from making meaning of statistical concepts may help to orient students (and their teachers!) towards this perspective. Describing a distribution as "more clumped in the center" conveys a more distribution-oriented perspective in language than stating, say, standard deviation or range to compare its dispersion.

The results of this study are not meant to downplay the importance of using conventional terms and measures in comparing groups. On the contrary, these are very important tools. Our hope is that teachers do not emphasize simply summarizing or reducing the data with conventional measures to make overly simplistic comparisons as we have seen schools do in examining test data (Confrey & Makar, 2005), but rather seek insights into the context the data represent through richer views that include notions of distribution and variation. Further research is needed to gain insight into *how* teachers understand concepts of variation and distribution, as well as to document how teachers support their students' emerging statistical understanding.

ACKNOWLEDGEMENTS

This research was funded by the National (U.S.) Science Foundation under their Collaborative for Excellence in Teacher Preparation program (DUE-9953187). The authors are grateful to NSF for its financial support and to the UTeach program at the University of Texas for their support in this study. The authors wish to thank the organizers of SRTL-3 for their permission and encouragement to make these results available to a broader audience, and also appreciated the feedback offered by our SRTL colleagues. The authors would also like to thank the reviewers who provided constructive, thoughtful, and extensive feedback on earlier drafts of this paper. Insightful and tenacious feedback from Iddo Gal, going beyond his role as editor of SERJ, greatly improved the final version.

REFERENCES

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools.* Published Doctoral dissertation. Utrecht, The Netherlands: Freudenthal Institute. [Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php]

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Biehler, R. (1997). Student's difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 169-190). Voorburg, The Netherlands: International Statistics Institute.

Boaler, J. (1997). *Experiencing school mathematics: Teaching styles, sex, and setting*. Buckingham: Open University Press.

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, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5-43.

Confrey, J. (1998). Voice and perspective: Hearing epistemological innovation in students' words. In M. Larochelle, N. Bednarz, & J. Garrison (Eds.), *Constructivism and education* (pp. 104-120). New York, N.Y.: Cambridge University Press.

Confrey, J., & Makar, K. (2002). Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA)*, *Athens, GA,* (Vol. 3, pp. 1267-1279). Columbus, OH: ERIC Clearinghouse, Mathematics and Environment Education.

Confrey, J., & Makar, K. (2005). Critiquing and improving the use of data from high-stakes tests with the aid of dynamic statistics software. In C. Dede, J. Honan, & L. Peters (Eds.), *Scaling up for success: Lessons from technology-based educational improvement* (pp. 198-226). Cambridge, MA: Harvard College of Education.

Confrey, J., Makar, K., & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *International Reviews on Mathematical Education (Zentralblatt für Didaktik der Mathematik), 36*(1), 32-40.

Cramer, D. (1997). *Basic statistics for social research*. London: Routledge.

Finzer, W. (2001). *Fathom*! (Version 1.16) [Computer Software]. Emeryville, CA: Key Curriculum Press.

Fischbein, E. (1987). *Intuition in mathematics and science: An educational approach.* Dordrecht, The Netherlands: D. Reidel Publishing Company.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

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/serj]

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

Hawkins, A. (Ed.). (1990). *Teaching teachers to teach statistics.* Voorburg, The Netherlands: International Statistics Institute.

Konold, C. (2002). "Hat Plots? " Unpublished manuscript, University of Massachusetts, Amherst.

Konold, C., & Higgins, T. (2002). Highlights of related research. In S. J. Russell, D. Schifter, & V. Bastable, *Developing mathematical ideas: Working with data*, (pp. 165-201). Parsippany, NJ: Dale Seymour Publications.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (submitted). Data seen through different lenses. *Journal for Research in Mathematics Education*.

Konold, C., & Miller, C. (2004). *Tinkerplots* (Version 1.0) [Computer Software]. Emeryville, CA: Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: research of critical barriers. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 151-167). Voorburg, The Netherlands: International Statistics Institute.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Kosslyn, S. (1994). *Elements of graph design*. New York: W. H. Freeman and Company.

Lemke, J. (1990). *Talking science: Language, learning, and values.* Norwood, NJ: Ablex Publishing.

Makar, K. (2004). *Developing statistical inquiry: Prospective secondary mathematics and science teachers' investigations of equity and fairness through analysis of accountability data.* Unpublished Doctoral dissertation, University of Texas at Austin (USA). [Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php]

Makar, K., & Confrey, J. (submitted). Modeling fairness in student achievement in mathematics using statistical software by preservice secondary teachers*. ICMI Study Volume 14: Applications and modeling in mathematics education*.

Makar, K., & Confrey, J. (2004). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353-373). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Marshall, J., Makar, K., & Kazak, S. (2002). Young urban students' conceptions of data uses, representation, and analysis. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA), Athens, GA*, (Vol. 3, pp. 1293-1304). Columbus, OH: ERIC Clearinghouse, Mathematics and Environment Education.

McClain, K. (2002). Supporting teachers' understanding of statistical data analysis: Learning trajectories as tools for change. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Moore, D. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants* (pp. 95-137). Washington, D.C.: Academy Press.

Moore, D., & McCabe, G. (1993). *Introduction to the practice of statistics* (2nd ed.). New York: Freeman.

National Research Council (1999). *High stakes: Testing for tracking, promotion, and graduation*. Heubert, J. P., & Hauser, R. M. (Eds.), Washington, D.C.: National Academy Press.

National Research Council (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, D.C.: National Academy Press.

National Research Council (2001). *Knowing what students know.* Washington, D.C.: National Academy Press.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M., & Begg, A. (2004). The school statistics curriculum: Statistics and probability education literature review. *Mathematics Curriculum Reference Group Report to the Ministry of Education.* Auckland, New Zealand.

Pfannkuch, M., Budgett, S., Parsonage, R., & Horring, J. (2004). Comparison of data plots: Building a pedagogical framework. Paper presented at the *Tenth International Congress on Mathematics Education (ICME-10)*, Copenhagen, Denmark, 4-11 July.

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

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Russell, S. J., & Mokros, J. R. (1990). What's typical: Teacher's descriptions of data. Paper presented at the *Annual Meeting of the American Educational Research Association*, Boston, MA.

Schoenfeld, A. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 311-343). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Shaughnessy, J. M. (2001). *Draft of grant proposal sent to the National Science Foundation*.

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

von Glasersfeld, E. (Ed.) (1991). *Radical constructivism in mathematics education.* Dordrecht, the Netherlands: Kluwer Academic Publisher.

Watkins, A. E., Scheaffer, R., & Cobb, G. (2003). *Statistics in action: Understanding a world of data*. Emeryville, CA: Key Curriculum Press.

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

KATIE MAKAR
School of Education, Social Sciences Bldg
University of Queensland
Brisbane, Queensland 4072
Australia

EXPLORING STUDENTS' CONCEPTIONS OF THE STANDARD DEVIATION

ROBERT DELMAS
University of Minnesota
delma001@umn.edu

YAN LIU
Vanderbilt University
yan.liu@vanderbilt.edu

SUMMARY

This study investigated introductory statistics students' conceptual understanding of the standard deviation. A computer environment was designed to promote students' ability to coordinate characteristics of variation of values about the mean with the size of the standard deviation as a measure of that variation. Twelve students participated in an interview divided into two primary phases, an exploration phase where students rearranged histogram bars to produce the largest and smallest standard deviation, and a testing phase where students compared the sizes of the standard deviation of two distributions. Analysis of data revealed conceptions and strategies that students used to construct their arrangements and make comparisons. In general, students moved from simple, one-dimensional understandings of the standard deviation that did not consider variation about the mean to more mean-centered conceptualizations that coordinated the effects of frequency (density) and deviation from the mean. Discussions of the results and implications for instruction and further research are presented.

Keywords: Standard deviation; Variability; Conceptions; Strategies; Interviews

1. INTRODUCTION

1.1. STUDENTS' UNDERSTANDING OF VARIABILITY

The current research study is partially motivated by the first author's collaborative research with Joan Garfield and Beth Chance on students' understanding of sampling distributions. One of the main findings from a decade of research is that this is a very difficult concept for students to develop (see Chance, delMas, & Garfield, 2004; delMas, Garfield, & Chance, 2004). DelMas, Garfield, and Chance (2004; Chance, delMas, & Garfield, 2004) speculate that the difficulty is partially due to students not having firm understandings of prerequisite statistical concepts such as center, distribution, and variability. With respect to variability, part of the difficulty may also stem from students' misunderstanding of how variability can be represented graphically. For example, when presented with a histogram, some students judge the variability of the distribution on the basis of variation in the heights of bars, or the perceived "bumpiness" of the graph, rather than the relative density of the data around the mean (Garfield, delMas, & Chance, 1999). Helping students develop a better understanding of variability and its representation may be one way to support a better understanding of sampling distributions.

Very little research has been conducted on students' understanding of variability (Reading & Shaughnessy, 2004; Shaughnessy, 1997), despite the central role the concept plays in statistics (Hoerl & Snee, 2001; Moore 1990; Snee, 1990) and an apparent conceptual gap in students' understanding of

Statistics Education Research Journal, 4(1), 55-82, http://www.stat.auckland.ac.nz/serj

variability (Reading & Shaughnessy, 2004; Shaughnessy, 1997). A few investigations have been conducted into students' understanding of sampling variability and instructional approaches that affect this understanding. Reading and Shaughnessy (2004) present evidence of different levels of sophistication in elementary and secondary students' reasoning about sample variation. Meletiou-Mavrotheris and Lee (2002) found that an instructional design that emphasized statistical variation and statistical process produced a better understanding of the standard deviation, among other concepts, in a group of undergraduates. Students in the study saw the standard deviation as a measure of spread that represented a type of average deviation from the mean. They were also better at taking both center and spread into account when reasoning about sampling variation in comparison to findings from earlier studies (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999).

Little is known, however, about students' understanding of *measures* of variation, how this understanding develops, or how students might apply their understanding to make comparisons of variation between two or more distributions. The latter ability represents an important aspect of statistical literacy that is needed both for interpreting research results and for everyday decision making. An understanding of statistical variation and of measures of variation is also needed to understand conceptually complex concepts such as sampling distributions, inference, and p-values (Chance, delMas, & Garfield, 2004; delMas, Garfield, & Chance, 1999; Saldahna & Thompson, 2002; Thompson, Saldahna, & Liu, 2004). An incomplete understanding of the standard deviation may limit students' understanding of these more advanced topics.

1.2. A CONCEPTUAL ANALYSIS OF STANDARD DEVIATION

Shaughnessy (1997; Reading & Shaughnessy, 2004) noted that the standard deviation is both computationally complex and difficult to motivate as a measure of variability. Part of this difficulty may stem from a lack of accessible models and metaphors for students' conceptions of the standard deviation (Reading & Shaughnessy, 2004). Most instruction on the standard deviation tends to emphasize teaching a formula, practice with performing calculations, and tying the standard deviation to the empirical rule of the normal distribution. This emphasis on calculations and procedures does not necessarily promote a conceptual understanding of standard deviation. A conceptual model of the standard deviation is needed to develop instruction that promotes the concept. We conjecture that such a model involves the coordination of several underlying statistical concepts from which the concept of standard deviation is constructed.

One of these fundamental concepts is distribution. The students in the current study worked with distributions of discrete variables, so the concept of distribution in this paper will be described in those terms. Essentially, an understanding of distribution requires a conception of a variable and the accumulating frequency of its possible values. Therefore, a visual or graphical understanding of distribution involves the coordination of values and density.

A second fundamental concept is that of the arithmetic mean. A conceptual understanding of the standard deviation requires more than the knowledge of a procedure for calculating the mean, either in procedural or symbolic form (e.g., $\Sigma x/n$). Imagery that metaphorically considers the mean to behave like a self-adjusting fulcrum on a balance comes close to the necessary conception. Such imagery supports the development of the third foundational concept, deviation from the mean. It is through the coordination of distribution (as represented by the coordination of value and frequency) and deviation (as distance from the mean) that a dynamic conception of the standard deviation is derived as the relative density of values about the mean. A student who possesses this understanding can anticipate how the possible values of a variable and their respective frequencies will, independently and jointly, affect the standard deviation. An illustration from the computer program used in the present study may help to clarify this conception.

To clarify the above ideas, Figure 1a presents screen displays from a computer program developed by the first author to study and promote students' understanding of the standard deviation. The program represents the distribution of a variable along the number line as a histogram made up of bars composed of a certain number of rectangles, each of which represents one data point (or observation). The location of a bar indicates the value of all the data represented by the bar (e.g., the tallest bar has eight data points each with a value of 4). The point location of the mean is indicated by an arrow along the number line. The standard deviation is reported as a numerical value below the arrow, and its size is represented by the length of a horizontal bar that extends below and above the mean. The deviation from the mean of each data point is printed within each rectangle. Overall, the program displays information that has the potential to enable students to mentally coordinate how changes in data points simultaneously affect the mean, deviations from the mean, and the standard deviation.

For example, a student can be asked to anticipate how the mean, deviations, and standard deviation shown in Fig. 1a are affected if the two lowest data points had a value of 1 instead of 2, resulting in moving the lowest bar on the left from 2 to 1 (see Fig. 1b). A student with a fully coordinated conception of standard deviation should anticipate that moving the lowest bar to a value of 1 shifts the mean to a slightly lower value and that all deviations from the mean would change simultaneously, i.e., the deviations of the two tallest bars increase, the deviations of the third tallest bar decrease, and possibly the deviations of the shortest bar increase. A student who is able to coordinate density (or frequency) with deviation should realize that the larger frequencies of the two tallest bars coupled with increases in deviation are likely to outweigh the few values in the third tallest bar that had a slight decrease in deviation. This would result in a larger density of values farther away from the mean and, therefore, an increase in the standard deviation.

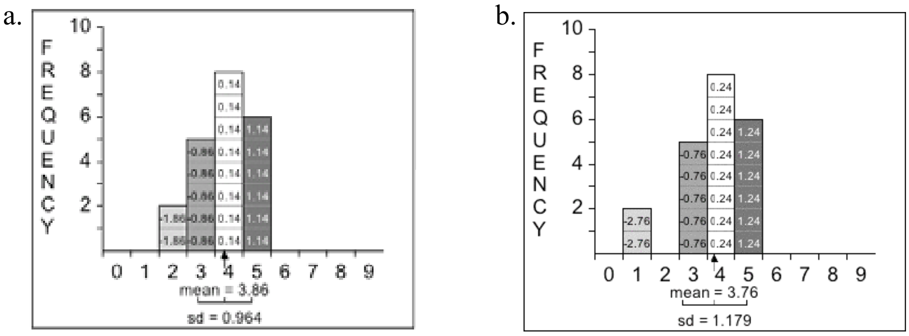


Figure 1. A graphic representation of standard deviation and its related concepts.

A student who understands these relationships should be able to reliably compare two distributions with respect to their standard deviations. Consider the two pairs of graphs presented in Figure 2a and 2b. Knowing only the location of the mean in both distributions, a student might reason that the graph on the left in Figure 2a has a larger standard deviation because there is a lower density of values around the mean. The same conception would lead to the conclusion that the graph on the right in Figure 2b has the larger standard deviation.

1.3. GOALS AND APPROACH

The main research goal of the current study was to gain a better understanding of how students' understandings of the standard deviation develop as they interact with a designed computer environment and an interviewer. The conceptual analysis provided in the previous section served as a framework for describing and interpreting students' understanding. The investigation was exploratory in nature and did not attempt to fully control factors that may or may not contribute to students' understanding of the standard deviation. Nonetheless, careful planning and consideration went into the design of the interactive experience so that it had the potential to promote the goal of a fully coordinated understanding of the standard deviation.

a.
b.

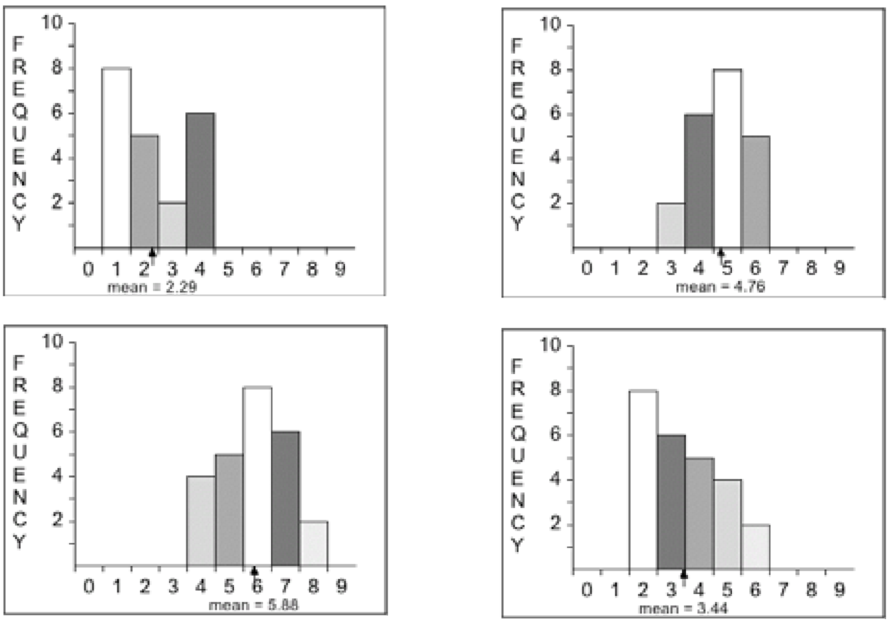


Figure 2. Comparing standard deviations in pairs of graphs.

In order to capture students' understanding, a computer application was written in Java and compiled for the Macintosh computer. The program was based on design principles for creating conceptually enhanced software (delMas, 1997; Nickerson, 1995; Snir, Smith, & Grosslight, 1995). Such software facilitates the development of conceptual understanding by accommodating students' current levels of understanding, providing familiar models and representations, supporting exploration by promoting the frequent generation and testing of predictions, providing clear feedback, and drawing students' attention to aspects of a situation or problem that can be easily dismissed or not observed under normal conditions. Conceptually enhanced software provides some of the conditions that foster conceptual change in science education (Chin & Brewer, 1993; Posner, Strike, Hewson, & Gerzog, 1982; Roth & Anderson, 1988). Among these conditions, the most pertinent for the present study is the engagement of students in a task that has the potential to repeatedly produce credible data that is inconsistent with students' current understanding in order to support reflective change of the underlying misconceptions.

Based on a pilot study, the computer program and interview protocol were designed with several instructional goals in mind. The primary instructional goal was to help students develop an understanding of how deviation from the mean and frequency combined to determine the value of the standard deviation. This entails that students distinguish value from frequency, recognize each value's deviation from the mean, understand that a distribution and its mirror image have the same standard deviation, and understand that the value of the standard deviation is independent of where the distribution is centered. A second instructional goal was to promote an understanding of how the shape of a distribution is related to the size of the standard deviation (e.g., given the same set of bars, a unimodal, symmetric distribution tends to have a smaller standard deviation than a skewed distribution).

2. METHODS

2.1. PARTICIPANTS

Students were recruited from four sections of an introductory statistics course at a large Midwest research university in the United States at the beginning of the spring 2003 term. Three of the course sections were taught by an instructor with a masters degree in mathematics education and four years of experience teaching the course. The other section was taught by a Ph.D. student in mathematics education with an emphasis in statistics education who had one year of experience teaching the course. The first author made a brief presentation about the nature of the research study to each of the four course sections during the second week of the term. Students were offered a $20 gift certificate to the university bookstore as an incentive. Of the 129 registered students, 27 initially volunteered to participate in the study. However, only thirteen of the students scheduled and participated in an interview. One of these students was dropped from the analysis because his responses indicated that he did not understand the nature of the research task. The final set of twelve students consisted of five males and seven females. As illustrated in Table 1, the group of participants did not differ noticeably from the non-participants, with the possible exception that a higher percentage of the participants received a final grade of A in the introductory statistics course when compared to the non-participants, although the difference is not statistically significant.

Table 1. Comparison of statistics students who did and did not participate in the study.

	Participants		Non-Participants			
Variable	n	Mean	n	Mean	Test Statistic	p-Value
Mean ACT Composite Score	11	19.8	108	19.2	$t(117) = 0.70$	0.486
Mean ACT Mathematics Score	11	19.0	108	18.2	$t(117) = 0.95$	0.342
Mean High School %-tile Rank	9	55.3	101	53.5	$t(108) = 0.36$	0.719
Cumulative College GPA	12	2.82	115	2.90	$t(125) = 0.43$	0.668
Cumulative College Credits	12	31.5	117	39.2	$t(127) = 1.00$	0.317
Percent Female	12	58.3	117	68.4	$\chi^2(1) = 0.15$	0.701
Percent Caucasian	12	58.3	117	58.1	$\chi^2(1) = 0.00$	0.999
Percent Course Grade of A	12	58.3	117	30.8	$\chi^2(1) = 2.58$	0.108
Percent Course Grade A or B	12	64.6	117	61.6	$\chi^2(1) = 0.001$	0.971

2.2. PRIOR INSTRUCTION

By the time this study started during the fourth week of instruction, the course had covered distributions, methods for constructing graphs (stem-and-leaf plots, dot plots, histograms, and box plots), measures of center (mode, median, and mean), and measures of variability (range, interquartile range, and the standard deviation). With respect to the standard deviation, students had participated in an activity exploring factors that affect the size of the standard deviation. During this activity, students compared nine pairs of graphs to determine which one had a larger standard deviation. Students worked in groups of two or three, identified characteristics of the graphs thought to affect the size of the standard deviation, recorded their predictions, and received feedback. The goal of the activity was to help students see that the size of the standard deviation is related to how values are spread out and away from the mean, the only factor associated with the correct choice across all nine pairs. Therefore, before the interviews, all of the students had received considerable exposure to the standard deviation as a measure of variability.

2.3. RESEARCH MATERIALS AND PROCEDURES

Students interacted with the computer program during interviews conducted by the first author. A digital video camera was used to capture a student's utterances and actions. The computer program wrote data to a separate file to capture students' actions and choices. Each movement or choice was time stamped so that it could be coordinated with the digital video recording.

Each interview was conducted in three phases: introduction, exploration, and testing. The interface for the introduction phase was designed so that students could become familiar with the controls and information displayed in the window. The interface for the exploration phase was similar to that of the introduction and presented students with tasks that required them to meet specified criteria. The interface for the testing phase presented ten test items designed to assess students' ability to compare the sizes of the standard deviations of two distributions after completing the exploration phase. Some of the same information presented during the exploration phase was available to support students' decisions. The following sections present more detail about each phase.

Introduction Phase

The introduction phase was relatively short and took only a few minutes. Students were introduced to the computer environment by moving two bars respectively representing frequencies of 5 and 8 within the graphing area. The program instructed the student to turn specified buttons on and off to display different information (mean, standard deviation, deviations, squared deviations, and the standard deviation formula). The interviewer described each type of information. Students were also asked to move the two bars within the graphing area while observing how the various values changed. When students indicated they were ready, they clicked a button labeled NEXT to start the exploration phase.

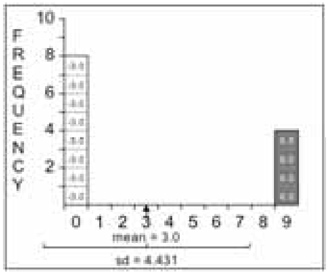
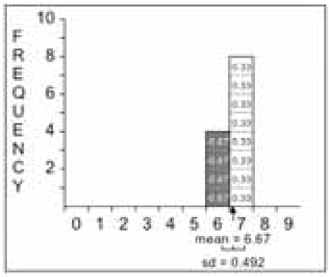
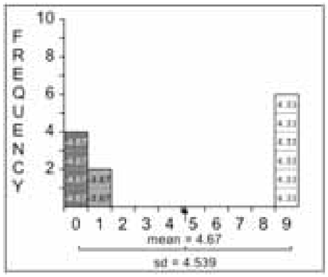
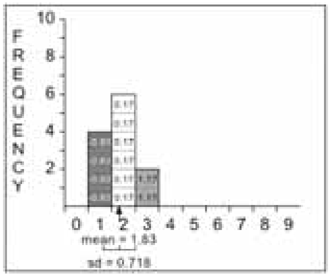
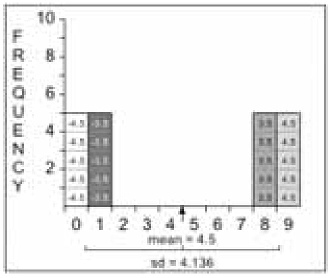
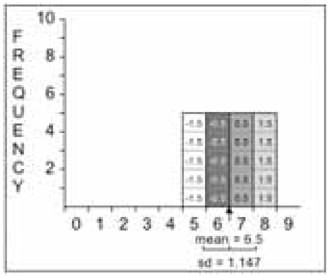
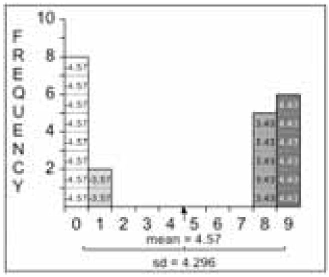
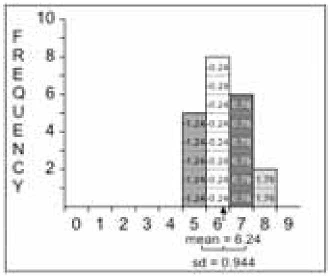
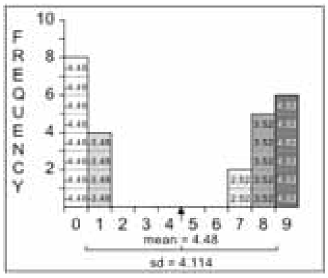
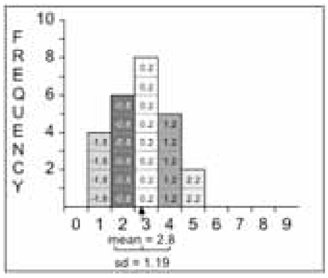
Exploration Phase

The exploration phase was designed to help students learn about factors that affect the standard deviation. Students could move frequency bars within the graphing area and observe simultaneous changes in the values of the mean, standard deviation, individual and summed deviations and squared deviations, and the standard deviation equation. Each student was presented with five different sets of bars (see Table 2), one set at a time. The number of bars ranged from two in the first set to five in the last. For each set of bars, the first task asked a student to find an arrangement that produced the largest possible value for the standard deviation, followed by a second task of finding another arrangement that produced the same value. The third through fifth tasks required three different arrangements that each produced the smallest value for the standard deviation. Each set of bars along with the five tasks comprised a "game".

Students used a CHECK button to determine if an arrangement met the stated criterion for a task. Students were asked to describe what they were thinking, doing, or attending to before checking. In addition, the interviewer asked students to state why they thought an arrangement did or did not meet a criterion once the CHECK button was selected and feedback received. The interviewer also posed questions or created additional bar arrangements to explore a misunderstanding or the stability of a student's reasoning.

Conceptual change can be facilitated through a combination of discovery learning and direct instruction (e.g., Burbules & Linn, 1988) and by drawing students' attention to relevant aspects that might be neglected (e.g., Hardiman, Pollatsek, & Well, 1986). When students had difficulty finding an arrangement that met a task criterion, or when they appeared to be moving in the direction of one of the goals, the interviewer used several approaches to support the student. A mean-centered conception of the standard deviation was promoted by drawing attention to the values of the deviations, by asking students how hypothetical movements of the bars would potentially affect the mean, and by the interviewer modeling reasoning of how the distribution of deviation densities affected the values of the mean and standard deviation. The second goal of promoting an understanding of how distribution shape is related to the size of the standard deviation was also supported. Students were asked to predict how the mean and standard deviation would be affected by bell-shaped and skewed arrangements of the same bar sets. Students were also asked to judge the extent to which a distribution was bell-shaped or whether the bars could be arranged to produce a more bell-shaped (or symmetric) distribution. If the student made a change to a distribution in response to a conjecture, the interviewer drew the students' attention to the value of the standard deviation and the direction of the change.

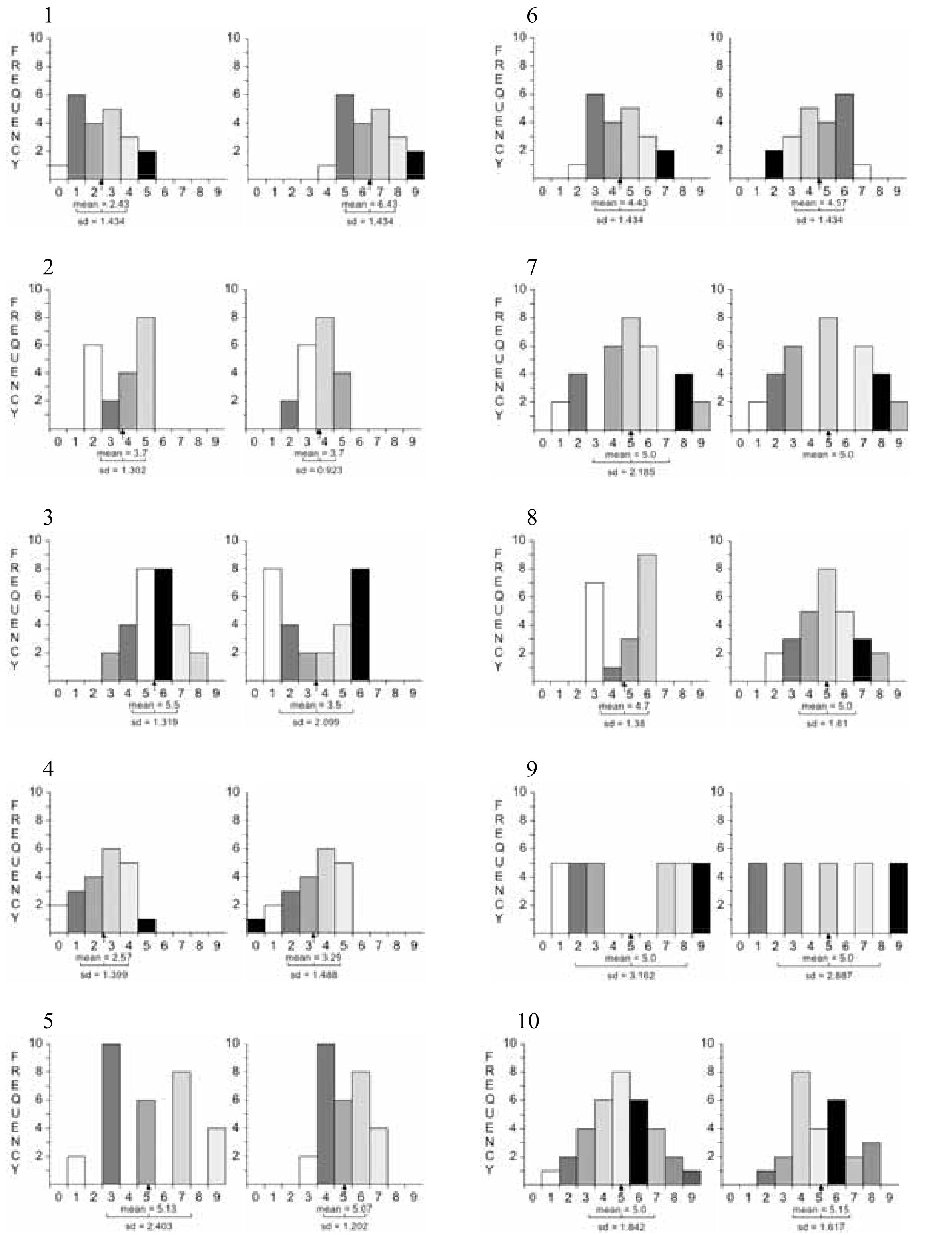
Table 2. Bar sets with possible solutions for the five games in the Exploration phase.

Game	Largest SD	Smallest SD
1		
2		
3		
4		
5		

Testing Phase

The testing phase was designed to assess students' understanding of factors that influence the standard deviation. Students were asked to answer 10 test items where each test item presented a pair of histograms (see Table 3). For each pair of histograms, the average and standard deviation were displayed for the graph on the left, but only the average was displayed for the graph on the right. The students were asked to make judgments on whether the standard deviation for the graph on the right was smaller than, larger than, or the same as the graph on the left. Once a student made a judgment, the investigator asked the student to explain his justification and reasoning. The student then clicked the CHECK button to check the answer, at which time the program displayed the standard deviation for the graph on the right and stated whether or not the answer was correct.

Table 3. Ten test items presented in the Testing phase.



The test items were based on the two stated goals of the study. None of the bar sets were identical to the sets of bars presented during the exploration phase. Each pair of graphs in test items 1 through 7 had identical bars placed in different arrangements. Test item 1 required students to recognize that both graphs had the same bar arrangement in different locations, whereas students had to recognize that one distribution was the mirror image of the other in test item 6. Test item 4 was specifically designed to see if students understood that given the same frequencies and range, a distribution with a stronger skew tended to have a larger standard deviation. Test items 2 and 3 tested students' sensitivity to the density of values around the mean. Test items 5 and 7 were similar in this respect, however the bars in both graphs were in the same order. These latter two items were designed to identify students who might attend only to the order of the bars and not to the relative density of deviations from the mean. The pair of graphs in test items 8, 9, and 10 did not have identical bars and also had a different number of values represented in each graph. In a pilot study, some students initially thought that evenly spacing the bars across the full range of the number scale would produce the largest value for the standard deviation. This misunderstanding was tested directly by test item 9, although this thinking could also be used to answer test item 5.

Test items 8 and 10 were designed to challenge the belief that a perfectly symmetric and bell-shaped distribution will always have a smaller standard deviation. Students were expected to find these items more difficult than the others. For both test items, each graph had characteristics that indicated it could have either the smaller or larger standard deviation of the pair. For example, in test item 8, the U-shaped graph on the left appeared to have less density about the mean while the graph on the right was perfectly symmetric with apparently a large portion of the density centered around the mean. Based on this, a student might have reasoned that the graph on the right had the smaller standard deviation. However, the graph on the left had a smaller range and represented a smaller number of values. Both of these characteristics made it possible for the graph on the left to have relatively more density around the mean than the graph on the right (which was the case). A reasonable response to test items 8 and 10 would have been that both standard deviations (or the variances) needed to be calculated to determine how they differ. The program provided the sum of squared deviations for both graphs to support these calculations.

3. RESULTS

3.1. EXPLORATION PHASE

The transcript analysis of the exploration phase focused on categorizing and describing the justifications students presented when asked why they expected an arrangement to satisfy a task criterion (e.g., the largest possible standard deviation). These justifications were taken to represent the students' conceptual understandings of the standard deviation as they developed during the interview. While some students appeared to start the interview with a fairly sophisticated understanding of factors that affect the standard deviation, most students started with a very simple, rule oriented approach.

Eleven broad categories of justifications were identified with the following labels (in alphabetical order): Balance, Bell-Shaped, Big Mean, Contiguous, Equally Spread Out, Far Away, Guess and Check, Location, Mean in the Middle, Mirror Image, and More Bars in the Middle. Each justification is described and illustrated in more detail in the subsequent sections. This section is organized in five subsections: Justifications for arrangements intended to produce the largest standard deviation (task 1) are presented first, followed by justifications for arrangements intended to produce the smallest standard deviation (task 3), and then justifications for arrangements intended to produce the same standard deviation (tasks 2, 4, and 5). These subsections are followed by a general discussion of students who began to coordinate conceptions about the standard deviation, and a final section that summarizes the findings from the exploration phase.

Graphs of students' bar arrangements are presented to illustrate excerpts from student interviews. Each bar in the software program was presented in a different color, however, the bars presented in the figures are in gray tones. Table 4 presents a legend for the bar colors.

Table 4. Legend of Bar Colors

□ = red	■ = blue	■ = orange	□ = green	□ = yellow

Largest Standard Deviation

For the first game, most students placed one bar at each of the extreme positions of the number line to produce the largest standard deviation. The subsequent games with more than two bars revealed more details of the students' thinking and strategies. In general, the arrangements represented placing the bars far apart or having them spread out. The typical order was to place the tallest bars at the extremes with subsequent bars placed so that heights decreased toward the center of the distribution. Three general strategies were identified based on students' justifications for their arrangements.

Far Away –Values

Some students stated that an arrangement should produce the largest possible standard deviation if the values (or the bars) are placed as far away from each other as possible. No mention was made of distance (or deviation) from the mean in these statements. This type of justification was prevalent across all the games.

Equally Spread Out

In the Equally Spread Out justification, the student believed that the largest standard deviation occurred when the bars were spread out across the entire range of the number line with equal spacing between the bars. This belief may have resulted from the in-class activity described earlier. Students may have translated "spread out away from the mean" to mean "equally spread out". For example, with three bars, the student places one bar above a value of 0, one above 9, and then places the third bar above 4 or 5. Some students using this approach realized that the third bar could not be placed with equal distance between the lowest and highest positions, so they would move the third bar back and forth between values of 4 and 5 to see which placement produced the larger standard deviation.

Nancy provided an example of this expectation. She first created the arrangement in Figure 3a when asked to produce the largest standard deviation in Game 2, followed by the arrangement in Figure 3b. The interviewer then asked her about the arrangements.

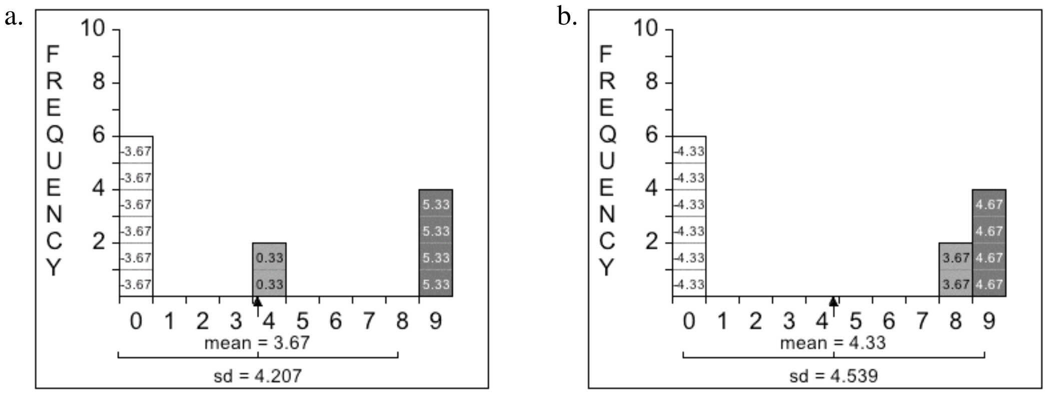


Figure 3. Nancy's arrangements for the largest standard deviation in Game 2.

Intv: And can you tell me what you were thinking, or trying out there.

Nancy: Well, at first I had the biggest bar…I just put the biggest bar because it was the first one on there. And I just put it somewhere. And then, at first I was just going to put the blue bar and then the orange bar, um, as far as, you know, just have them kind of equally spread out. But then I thought, maybe if I got the orange bar closer to the mean, there's less data on the orange bar, so, er, um, yeah, there's less data there.

Intv: There's fewer values.

Nancy: Fewer values, yeah. So, I wanted to put that closer to the mean than the blue bar because the blue bar has more, more data.

Intv: Mm hmm.

Nancy: So, and then, I didn't know if I put the orange bar right on top of the mean, or if I put it over here by the blue bar, how much of a difference that would make. But I figured the more spread out it would be the bigger the deviation would be, so.

Intv: Okay, so by having it kind of evenly spaced here, is that what you're thinking as being more spread out?

Nancy: Yeah. Yeah. Having, like, as much space as I can get in between the values.

Far Away –Mean

This rule is similar to Far Away-Values, except that the largest standard deviation is obtained by placing the values as far away from the mean as possible. Carl provided a clear expression of this rule in the first game, although he was not sure that it was correct.

Intv: Now, before you check, can you tell a little about what you were thinking as you were trying things out there?

Carl: Well…um…what I know about standard deviation is, um, given a mean, the more numbers away, the more numbers as far away as possible from the mean, is that increases the standard deviation. So, I tried to make the mean, um, in between both of them, and then have them as far away from the mean as possible. But I don't, yeah, I don't know if that's right or not, but, yeah.

Later, in Game 2, Carl was trying to arrange three bars to produce the largest possible standard deviation. He first used an Equally Spread Out approach (see Figure 4a) and then appeared frustrated that it did not produce the intended result. The interviewer intervened, trying to help him extend the Far Away – Mean rule he had used earlier to the new situation. With this support, Carl successfully produced the distribution in Figure 4b. The interviewer then asked Carl to explain why he thought the arrangement resulted in the largest standard deviation.

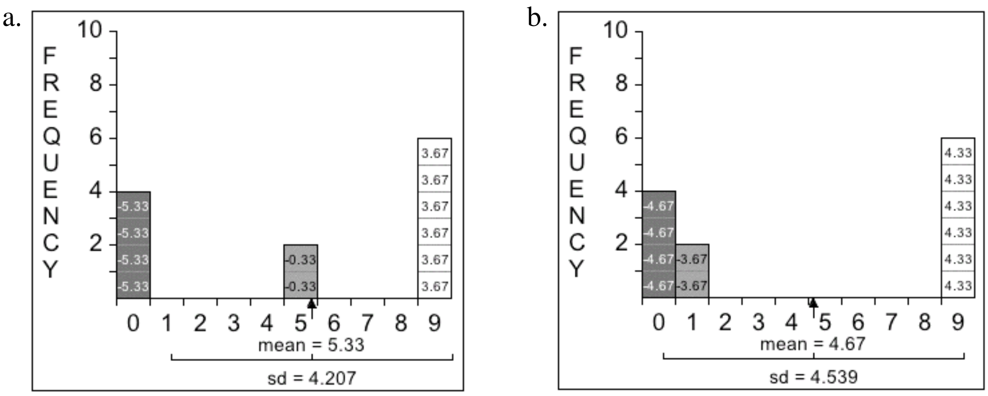


Figure 4. Carl's arrangements for the largest standard deviation in Game 2.

Carl: Because…basically because I have the most amount of numbers possible away from the mean. Um, because if I were…say I have two bars that are bigger than the one smaller bar,

Intv: Mm hmm.

Carl: If I were to switch these two around.

Intv: The blue and the orange one?

Carl: The blue and the orange. The standard deviation would be smaller then because there's more of the numbers closer to the mean. So, yeah.

Balance and Mean in the Middle

With the Balance strategy, the bars are arranged so that near equal amounts of values are placed above and below the mean. Typically, the bars were arranged in a recognizable order; descending inward from the opposite extremes of the number scale to produce the largest standard deviation. This justification started to emerge during Game 2 with three bars of unequal frequency. Mary provided an example of a Balance justification during Game 2.

Mary: Because, and just with a frequency bar of six over at the high-end, um, it makes it so that the mean can stay more towards the middle. Because if you put too much frequency at one end it kind of makes the mean, the mean kind of follows it. Or, I, so I think that this kind of makes it, evens it out so that it, it can be, um, more frequency away from the mean to get the highest standard deviation.

A Balance justification was often combined with another justification, such as Far Away-Value or Far Away-Mean, during the later games. Another justification that tended to occur with a Balance statement was the Mean in the Middle justification. Alice provided an example where both Mean in the Middle and Balance were used to justify an arrangement (see Figure 5) for producing the largest standard deviation in Game 4.

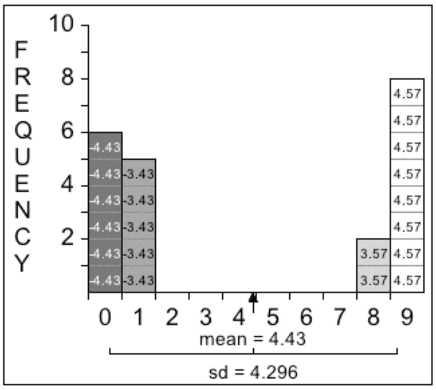


Figure 5. Alice's arrangement for the largest standard deviation in Game 4.

Alice: OK. I think this one will.

Intv: And can you tell me why?

Alice: Because, again I wanted to kind of keep the same values on both ends to kind of get the mean to fall in the center.

Intv: Mm hmm.

Alice: And then, that way they'd both be as far away from the mean as possible. And I think if you put the two middle bars together, then the two, like the largest and the smallest one, you'll get about the same value.

Intv: About the same number of values at each end?

Alice: Yeah, the same number of frequencies.

Big Mean

Troy was the only student who initially believed that a distribution with a higher mean would have a larger standard deviation. The first statement of this belief came in Game 1 after finding an arrangement that produced the largest standard deviation. He immediately posed the question "What would it be if I switched the bars around?" When asked what he thought would happen, he stated "I was thinking that if the mean was, I knew the mean was going to be smaller. I thought that would make the standard deviation smaller."

Troy's reasoning seemed to be primarily one of association: the mean gets lower, so the standard deviation will become lower. While he appeared to have an understanding of how value and frequency affect the mean, he had not coordinated these changes with changes in the standard deviation. Even though Troy witnessed that the standard deviation did not change when he produced a mirror image of the distribution, his belief that the size of the mean affects the standard deviation persisted. The following excerpt is from Game 1 after Troy found a distribution with the smallest standard deviation.

Troy: So I suppose it wouldn't, again, it wouldn't matter if I, like the other graph, if I moved them both together to a different, uh, number on the x-axis, right?

Intv: Yep, well, you're going to get a chance to find out.

Troy: OK. Again. Move at least one bar.

Troy: Check?

Intv: Mm hmm.

Troy: Yeah, it's the same.

Intv: Mm hmm. So does that surprise you, or, that the standard deviation is the same?

Troy: I guess a little bit. I, I don't know. I just, I guess, again, I thought if you. Again I guess it's thinking that if the mean goes up that the standard deviation goes up. But obviously that's not true a case.

Troy seemed to have learned that the location of an arrangement did not affect the standard deviation, although he did not appear to understand why at this point. He did not, however, present a Big Mean justification during any of the remaining games.

Smallest Standard Deviation

Nearly all of the students placed the bars next to each other in some arrangement to produce the smallest possible standard deviation. Several distinct variations were identified.

Contiguous

Students using this approach stated that they placed the bars next to each other, or as close together as possible, to obtain the smallest standard deviation. In addition to the statement of contiguity, three-fourths of the students placed the bars in an ascending or descending order based on bar height. Alice used a Contiguous – Ascending Order (see Figure 6) in her first attempt to come up with the smallest standard deviation for the three bar situation in Game 2.

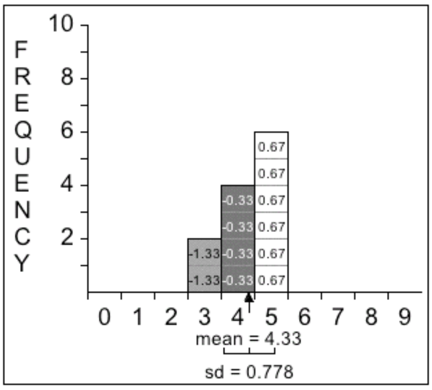


Figure 6. Alice's arrangement for the smallest standard deviation in Game 2.

Intv: Can you tell me about, uh, what you were doing there?

Alice: Um, well I knew I wanted the bars to be right next to each other.

Intv: Mm hmm.

Alice: Um, and I wanted the mean to kind of fall somewhere in the blue part right here, the middle one.

Intv: OK

Alice: And, so I was just going to check to make sure that if I put the smaller one in the middle that I didn't change, like, the deviation would be smaller or larger.

Intv: OK. And I noticed that the bars are going from the shortest to the next tallest to the tallest.

Alice: Yep.

Intv: And any reason why you have put them in that order?

Alice: Um, well I was thinking, like I think of, um, numbers. And if I count like the boxes there is one, two, three, four, five here, four there, and two there.

Intv: Mm hmm.

Alice: I kind of calculate where the median would be and then make sure that the numbers kind of fall somewhere in the middle. So.

Intv: OK. You want to check this one out?

Alice: OK

Intv: So it says it can be even a little smaller.

Alice's reference to the median is not clear, but she may have identified the second tallest bar as the bar with the "median" height, and this was the motivation for placing it in the middle. Alice found out that this arrangement was not optimal and quickly produced an arrangement that did result in the smallest standard deviation.

All of the students who produced an ascending or descending order in Game 2 appeared to abandon this approach for the later games, with one exception. Mona was the only student to use a Contiguous-Descending Order in Game 4 (see Figure 7), even though she produced a bell-shaped distribution in Game 2 for the smallest standard deviation. The following excerpt illustrates her thinking about why the arrangement will produce the smallest standard deviation.

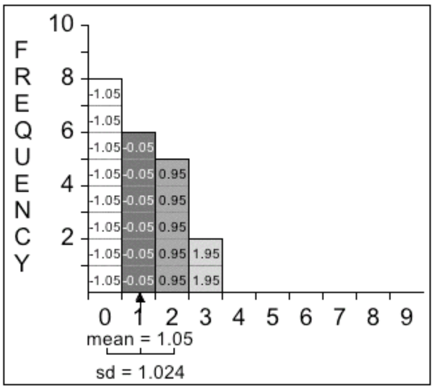


Figure 7. Mona's arrangement for smallest standard deviation in Game 4.

Mona: Yeah, I think this one is right.

Intv: OK. Do you want to check?

Mona: Wait, no, it's not right.

Intv: Now why were you thinking it was right, and why are you thinking it's not right?

Mona: Wait, OK, maybe it is right. OK, because um, I know to make it like, to make it, the standard deviation lower, it has to be closer to the mean. And a lot of them have to be closer to the mean. So, this one is right by the mean almost.

Intv: The red [tallest] bar?

Mona: Yeah.

Intv: Mm hmm

Mona: So it's like, the more of these I have closer to it the better it is. And, and since I only have two that's not, then...

Intv: Yeah. I see what you're saying. You want to check it out?

Mona: OK

Intv: So now it says you can make it smaller.

Mona's statements indicate an understanding that most of the values need to be close to the mean in order for the standard deviation to be small. Her statements indicate that she was considering the placement of values in relationship to the mean, but not necessarily considering deviations from the mean, or the relative density of the deviations. Her thinking ("a lot of them have to be closer to the mean") may have been too general to guide an optimal solution. Mona did eventually produce a bell-shaped distribution with the smallest standard deviation, and she did not use an ascending or descending order in Game 5.

More Bars in the Middle

Some students stated that one of the reasons the standard deviation would be the smallest was because more values or the tallest bars were placed in the middle of the distribution or close to the mean. This justification started to appear in Game 2. A More Bars in the Middle statement, coupled with a statement of contiguity, was the predominant justification offered in Game 4. Mona provided a More Bars in the Middle statement in the previous excerpt when she stated, "a lot of them have to be closer to the mean." Carl used this justification for his first arrangement in Game 2 (see Figure 8).

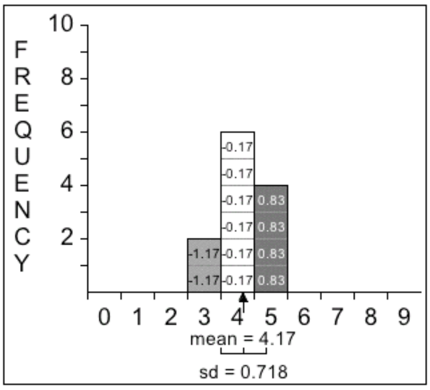


Figure 8. Carl's first arrangement for the smallest standard deviation in Game 2.

Intv: So what were you doing here?

Carl: I put…I tried to scrunch up all of the numbers as close as I could to the mean, and tried to make a new mean in this case. And, uh, it wouldn't have worked if I were to have switch, if I were to have it in progressive order of small orange, medium blue, to big red.

Intv: Mm hmm.

Carl: Because the red is the biggest. You want it to be that in between so the mean is going to, so that you'll have the mean more with the more amount of numbers, basically.

Nancy also used a More in The Middle justification for her first smallest standard deviation arrangement in Game 4. She started off with an arrangement that was near optimal (see Figure 9a) and then decided to switch the positions of the two shortest bars to produce the graph in Figure 9b. Although the result was fairly symmetric and bell-shaped, she talked only about placing lots of values close to the mean.

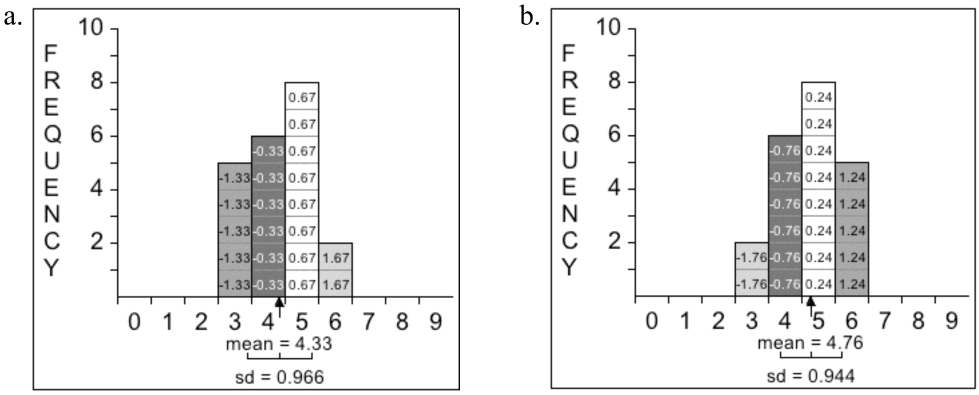


Figure 9. Nancy's arrangements for the smallest standard deviation in Game 4.

Nancy: Yeah. Oh. I've got all of the biggest values together so, there…it was like the one with three where I had them in order. Because I had them kind of in order, and then I had the smallest one. But now I'm just getting all the biggest values closest to the mean. Okay. Before I didn't, so.

The combination of Contiguous with More Bars in the Middle appears to represent a more complex understanding of factors that affect the standard deviation where the students moved from considering only the closeness of the values in relation to each other to also considering the relative density about a perceived center of the distribution. While not as complete as the completely coordinated conception of the standard deviation outlined earlier, this thinking is a closer approximation than most students demonstrated during the first two games.

Bell-Shaped

Students who made a Bell-Shaped justification stated that they were trying to produce a symmetrical or bell-shaped arrangement. Except for Game 3, none of the bar sets allowed the creation of a perfectly symmetrical distribution. This type of statement was the primary justification offered in Game 5, being made by only one student during Game 2 and two students during Game 4. This is probably the result of the interviewer drawing students' attention to the bell shape of distributions that produced the smallest standard deviation in Game 4.

Students who offered Bell-Shaped justifications typically positioned the tallest bar first, perhaps to function as an anchor or central point. The other bars were placed to the left and right of this central location. A few students were methodical in their placement, alternating the other bars to the left and right of center in order of height. Others seemed to place a few of the tallest bars to the left and right of center, and then checked the size of the standard deviation as they tested out various arrangements.

Mona appeared to use a Bell-Shaped approach in Game 5 when trying to come up with the smallest standard deviation, but she did not have a method that initially produced the most symmetric or bell-shaped arrangement of the five bars (see Figure 10). She started with the arrangement in Figure 10a, saying that "it has to have sort of like a normal, normal shape, and this kind of looks normal to me." The interviewer intervened by asking, "Is that as normal as you can make it look? Is that as bell shaped as you can make it look?" Mona then switched the locations of the two shortest bars to produce the graph in Figure 10b, finding that it produced a smaller standard deviation, and judging that the arrangement was more bell shaped.

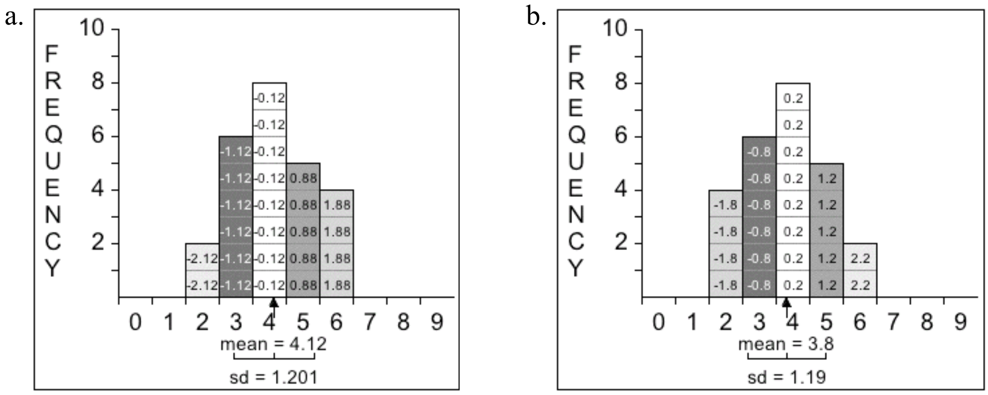


Figure 10. Mona's arrangements for the smallest standard deviation in Game 5.

Balance

The Balance justification was introduced earlier when describing statements students made about arrangements for the largest standard deviation. A balance approach was also used to create an arrangement that produces the smallest standard deviation, as illustrated by Carl in Game 4. Carl made the same type of arrangement as represented in Figure 9b. The interviewer asked why this would produce the smallest standard deviation.

Carl: Um, because, um, the way I have it right now, the red [tallest] bar's in the middle, the blue's [second tallest] on the left, the orange [third tallest] is on the right. And when I had the green [shortest] on the right, there was less numbers in the orange to throw off, um, the two extra numbers that were going to be there at the six value. So, by placing it on the other side, there's more – there's one more number with the blue bar to throw off one of the numbers in the green bar, so it will bring it a little bit closer.

Intv: Okay. When you say throw off, um, if I were to say balance?

Carl: Yeah.

Intv: Is that a similar thing to what you're saying.

Carl: Yeah. The balance of the scale. It's like having, um, I don't, um, It's kind of like trying to weigh on a scale, or like a teeter-totter. Um, there was – you have a lot in the middle, and some on the sides, but by placing it on one of the sides, it would actually teeter more that way because there's less values in the middle for that one. But if you place it on the other side, it would teeter less because there's more values in the middle to help, yeah, keep it in balance.

The Balance justification may be an extension of the Bell-Shaped justification in that the approach produces a somewhat symmetric and bell-shaped distribution. However, the Balance justification expresses a deeper understanding of the interrelationship between the frequencies of values, their relative placement with respect to the mean, and the overall effect on the standard deviation. The Bell Shape justification seems to be more rule-oriented, using a visual template to produce a distribution that is expected to minimize the standard deviation.

The Same Standard Deviation

Students used two general approaches to make arrangements that produced the same standard deviation.

Mirror Image

The student would often create a mirror image of a distribution to produce a distribution with the same standard deviation. Some students expressed this approach as "swapping" or "switching" the positions of the bars, while others stated that the arrangement was "flip-flopped". The student would state that since the bars were still as close together or as far apart as before (either relative to each other or relative to the mean), the standard deviation should be the same. Some students demonstrated that they understood that a mirror image would produce the same value for the standard deviation following a prompt from the interviewer. After Jane created the graph in Figure 11, the interviewer posed a situation for her to consider.

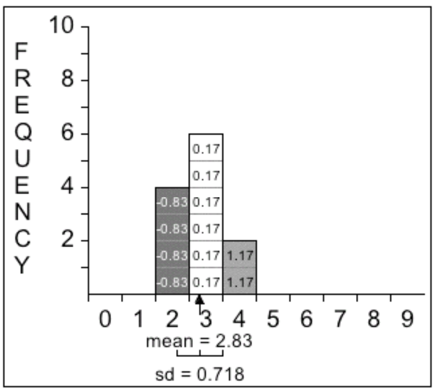


Figure 11. Jane's third arrangement for the smallest standard deviation in Game 2.

Intv: Um, I'm going to move the blue bar off to the side, put the orange bar on the left side of the red.

Jane: OK

Intv: And just ask you to leave those two there.

Jane: OK

Intv: Can you still come up with the smallest standard deviation?

Intv: OK. And what are you doing there?

Jane: I just moved the blue bar back next to the red one. It is just a mirror, it's one of those mirror imagy things again where you just switch the two. It doesn't matter which side, just as long as you have the biggest bar closer to the mean.

Location

In this approach, the student would move all the bars an equal distance, maintaining the same relative arrangement and standard deviation. The student would note that the standard deviation should be the same because the relative distances between the bars or from the mean were maintained. Troy made the graphs in Figure 12a and Figure 12b for the fourth and fifth tasks in Game 5. After checking and finding that the last graph did produce the smallest standard deviation, the interviewer followed with a question.

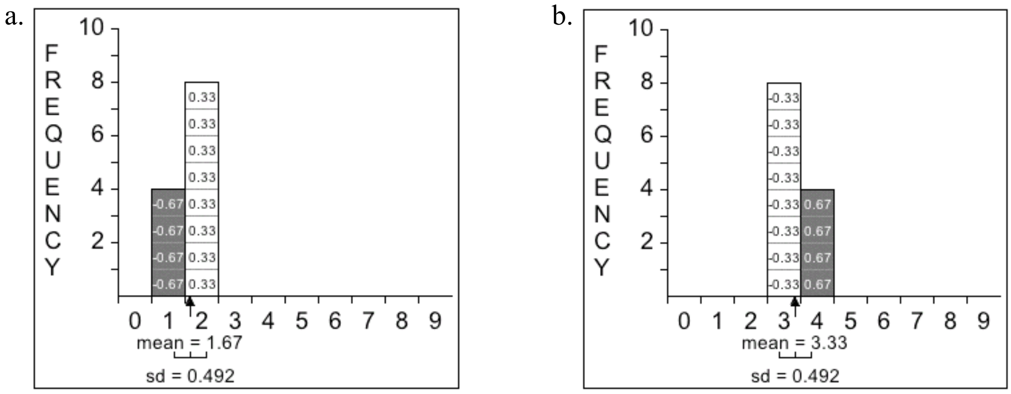


Figure 12. Two of Troy's arrangements for the smallest standard deviation in Game 1.

Intv: Can you tell me anything about what you think is going on now with, with the relationship of the bars to each other and the size of the standard deviation?

Troy: Oh, just that the closer the bars are together the lower the standard deviation is.

Intv: Mm hmm.

Troy: Vice versa.

Intv: OK

Troy: Um. And it doesn't matter where they are, and what the mean is or where they are on the x-axis.

Coordination of Concepts

Review of Carl's More in the Middle justifications during Game 2 indicates that he was able to coordinate the effects of the values and the frequencies represented by the bars on the value of the mean. Carl also appeared to display some coordination between the effects of frequency and deviations from the mean and the value of the standard deviation, although his focus was primarily on having most of the values (the tallest bar) close to the mean. None of the other students demonstrated this type of coordination during Game 2. Carl continued to demonstrate an increasing coordination of concepts as he progressed through the games.

Several more students mentioned deviations from the mean and demonstrated conceptual coordination in their explanations during Game 4. Mary started with the arrangement in Figure 13 for the task of producing the largest standard deviation. When asked why she thought the arrangement would produce the largest standard deviation, Mary offered a Far Away-Mean explanation:

Mary: I think we did this for another one where we made sure the highest frequencies were as far away from the mean, or as possible, and then try to just do it in order of the biggest to the smallest and keep them all grouped away. So I don't know, 4.175. I bet you it will maybe work this way too.

Mary was generalizing from Game 2 to Game 4, but did not have a full understanding of how to maximize the size of the deviations from the mean. However, she did express a coordination of both value and frequency with respect to distance from the mean and the relative size of the standard deviation.

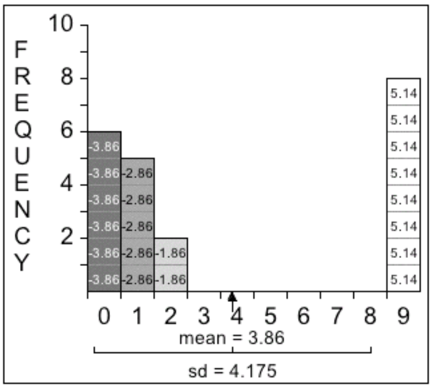


Figure 13. Mary's first arrangement for the largest standard deviation in Game 4.

Linda also demonstrated an understanding that increasing the number of large deviations increases the standard deviation.

Linda: Um, I put like the largest bars on the two outside scores and the smaller ones on the inside just so, like, the bigger chunks of them would be farther away from the mean and the smaller ones might, I didn't think that would affect the standard deviation as much as the larger values.

Adam, who first gave a Balance type of explanation, also observed that there was a pattern to how the bars were placed and connected this to deviations from the mean.

Adam: I'm guessing this smaller, the smaller, the smallest sets of data values have to be towards the mean. Has to be closer towards the mean than the larger sets, because if you want the standard deviation, which is the larger amount of numbers away from the mean, then you are going to have to take out the largest sets of numbers and keep them farther away from the mean and keep the smallest sets closer to the mean.

Here, Adam appears to coordinate the effects of frequency and deviation after the fact.

Alice demonstrated a more complex understanding of how frequency and deviation from the mean combined to affect the size of the standard deviation. As discussed earlier, Alice produced the graph in Figure 5 and justified why it produced the largest standard deviation with a combination of Balance and Far Away – Mean statements. The interviewer followed her statement with a hypothetical question:

Intv: If we were to switch positions of the gold bar [frequency of 5] and the green bar [frequency of 2], what do you think would happen to the standard deviation?

Alice: It would get pulled towards the gold and the red bar. Because there is more frequency over there than on the other two. Because they aren't pulling the mean that much more over.

Intv: OK. What do you think would happen to the standard deviation, then?

Alice: I think it would get smaller.

When Alice stated "it would get pulled towards the gold and the red bar," she appeared to be referring to the mean. As such, her prediction would be correct. She demonstrated an understanding that the mean would shift towards the larger cluster of values, creating a larger number of smaller deviations from the mean, and subsequently result in a smaller standard deviation. Carl and Alice were the only students to demonstrate this level of complexity in their justifications during Game 4.

Summary of Exploration Phase Justifications

Table 5 presents a list of the justifications that students gave for bar arrangements produced during the exploration phase. The entries in Table 5 indicate the number of instances of each justification across students. The tallies for the five games provide a sense of when a justification first appeared and if a justification continued to be used. The tallies illustrate distinctions among the justifications with respect to the type of tasks for which they were most likely to be used.

The first five justifications listed in Table 5 represent reasons given primarily for arrangements in the first task, ordered from most to least frequent. The next three justifications are those used primarily for the third task where the student was asked to produce an initial arrangement with the smallest standard deviation, again ordered from highest to lowest frequency. The Mirror Image and Location justifications represent reasons given for arrangements that produced the same standard deviation as a previous arrangement (tasks 2, 4, and 5). The Guess and Check approach is listed last because it was used with almost equal frequency in tasks 1 and 3, and represents a general strategy more so than a justification.

Table 5. Number of Instances of Each Justification Used by Students in the Exploration Phase.

						TASKS					
	GAMES					Largest SD		Smallest SD			
JUSTIFICATION	1	2	3	4	5	1	2	3	4	5	TOTAL
Far Away	12	8	12	8	5	43	2				45
Balance		7	8	10	10	30		4		1	35
Mean in the Middle	2	2	5	4		8	1	4			13
Equally Spread Out		7	1			8					8
Big Mean	1					1					1
Contiguous	17	11	12	7		1		42	2	2	47
More Bars in the Middle		6	1	9	3			18		1	19
Bell-Shaped	1	1		2	7	1		10			11
Mirror Image	28	27	23	24	22		59		35	30	124
Location	12	13	15	13	14			4	27	36	67
Guess and Check	6	15	3	14	9	25	1	20	1		47

The Equally Spread Out justification was used primarily during the second game with three bars, and then dropped out of use after feedback indicated that the approach did not produce a distribution with the largest standard deviation. The idiosyncratic Big Mean justification was only used once in the first game, dropping out of use after the student received feedback that it was not effective. Far Away and Balance became the predominant justifications for arrangements intended to produce the largest standard deviation. A statement that an arrangement tended to place the mean in the middle of the distribution accompanied Far Away and Balance statements for a few students, but was not used frequently.

Contiguity was the primary justification used for the first arrangement intended to produce the smallest standard deviation, although it was not used at all for the fifth game. Students also tended to accompany Contiguity justifications with a recognition that an arrangement placed more of the taller bars in the middle of the distribution. This coupling of Contiguity and More Bars in the Middle was predominant during the second game with three bars and the fourth game with four bars of unequal frequencies. The tallies also illustrate that the Mean in the Middle and the More Bars in the Middle justifications appear to be distinct in that the former occurred primarily for the largest standard deviation criterion, whereas the latter was used exclusively for the smallest standard deviation criterion. Statements that an arrangement produced a bell-shaped distribution were made primarily during the fifth game, presumably as a result of prompting from the interviewer.

The predominant strategy for producing a new arrangement with the same standard deviation as a previous arrangement was to create a mirror image of the prior distribution. As would be expected by the constraints of the second task, this was the only justification given for producing a second distribution with the largest standard deviation. The Mirror Image and the Location justifications were used with near equal frequencies for tasks 4 and 5 which required additional arrangements with the smallest possible standard deviation.

A Guess and Check approach was used primarily when producing the first arrangement for either the largest or smallest standard deviation criteria. Once an arrangement was found that met the criterion, the Mirror Image or Location justifications were used to produce subsequent arrangements that met the criterion.

3.2. TESTING PHASE

Table 6 presents the justifications students used in their responses to the 10 test items. Item 8 was the most difficult item with only one student making the correct choice. Nine of the 12 students answered 9 of the 10 test items correctly. The other three students answered 7 of the test items correctly. Given the high correct response rate, correct responses were not associated with any particular justification or combination of justifications.

Table 6. Number of Students Who Used Each Justification to Explain Choices During the Testing Phase.

TEST ITEM 1		TEST ITEM 2		TEST ITEM 3		TEST ITEM 4		TEST ITEM 5	
Location	12	More in Middle	9	More in Middle	6	Bell	3	Contiguous	7
		Bell Shape	3	Far Away	4	More in Middle	3	Mean in Middle	1
		Range	1	Mean	2	Balance	1	More in Middle	1
		Contiguous	1	Value	1			Equal Spread Out	1
				Bell Shape	2			Range	1
				Contiguous	1				
TEST ITEM 6		TEST ITEM 7		TEST ITEM 8		TEST ITEM 9		TEST ITEM 10	
Mirror Image	11	Contiguous	7	Calculation	7	More in Middle	6	Calculation	9
Mean in Middle	1	More in Middle	5	More in Middle	4	Mean in Middle	3	Range	4
		Far Away	5	Mean in Middle	4	Far Away	3	More in Middle	2
		Mean	5	Bell Shape	4	Mean	3	More Values	2
		Mean in Middle	1	More Values	4	Calculation	2	Mean in Middle	1
		Range	1	Range	2	Contiguous	2	Bell Shape	4
				Contiguous	2			Frequency	1

Students' justifications for their responses tended to be similar to those identified during the exploration phase, and tended to reflect relevant characteristics of the graphs. For example, all students indicated that the two graphs had the same standard deviation for the first test item, noting that the arrangements were the same but in different locations. Most of the students' justifications for test items 2 and 3 were More in the Middle statements, although test item 3 also prompted Far Away justifications, probably because the graph on the right was U-shaped. For test item 6, students recognized that one graph was the mirror image of the other and that the standard deviations were the same. Almost all of the students gave a Mirror Image justification for their choice in item 6.

Students found test item 4 a little more difficult than the first three tests. Most noted that the only difference between the two graphs was the placement of the black (shortest bar). Only about half of the students provided a justification for their responses, which tended to be either that the graph on the left was more bell-shaped, or that it had more values in the middle or around the mean compared to the graph on the right. A few students stated that the left-hand graph was more skewed.

Two students (Adam and Lora) judged incorrectly that the graph on the left would have a smaller standard deviation in test item 4. Adam noted the different location of the black bar in both graphs, but did not offer any other justification for his response. Adam did not demonstrate a consistent understanding of how frequency and deviation from the mean combined to affect the standard deviation during the exploration phase, which was evidenced by his extensive use of a Guess and Check approach. Item 4 may have proven difficult due to this lack of understanding. Lora also displayed Guess and Check behavior during the exploration phase, but not to the same extent as Adam. Lora's justification was actually a description of each graph's characteristics and did not compare the two graphs on features that would determine how the standard deviations differed. When asked how the graph on the right compared to the graph on the left, Lora responded, "That, well that's kind of the same thing as this one, but this one is, um, the smallest one is on the end instead, and along with the other one." No other justifications were offered before checking.

The difficulty with test item 4 may have resulted from the nature of the tasks during the exploration phase. The tasks required solutions that had some degree of symmetry and did not require students to produce skewed distributions, although skewed distributions were often generated and checked before an optimal distribution was found for the third task. The arrangements produced for the largest standard deviation were U-shaped, with a large gap through the middle, while the optimal arrangements for the smallest distribution were bell- or mound-shaped. This experience during the exploration phase probably facilitated the comparison in test items 2 and 3, but did not necessarily provide experience directly related to the comparison needed to solve test item 4. Nonetheless, the majority of the students correctly identified the graph on the right as having a larger standard deviation.

The first four test items involved graphs with contiguous bar placement, so that test item 5 provided the first test of sensitivity to gaps between the bars. Almost all of the students' justifications for test item 5 involved statements of contiguity. A few students noted that the range was wider in the left-hand graph so that the graph on the right should have a smaller standard deviation given that the bars were the same and in the same order in both graphs. One of these students noted that the bars were equally spread out for the graph on the left, although the emphasis was on the wider range. Two students, Adam and Jeff, appeared to ignore the gaps between bars in the left-hand graph and responded that the two graphs would have the same standard deviation, suggesting that shape was the main feature they considered in their decisions and an insensitivity to spread or deviation from the mean.

Test item 7 also tested students' understanding of how gaps affected the standard deviation, although in a more subtle way. All students answered test item 7 correctly, using arguments of contiguity and more values in the middle for the left-hand distribution as the reason why the right-hand distribution would have a larger standard deviation. Some students also noted that the right-hand distribution had relatively more values away from the mean. Test item 9 also presented gaps in the two distributions and challenged the misconception that having the bars evenly or equally spaced produces a large standard deviation. All students answered test item 9 correctly and offered justifications similar to those used in test item 7, indicating that the initial misconception represented by the Equally Spread Out justification had been overcome by those who first displayed it.

Test items 8 and 10 were designed to challenge the idea that perfectly symmetric, bell-shaped distributions always have the smaller standard deviation. For test item 8, all but one of the students incorrectly stated that the graph on the right would have a smaller standard deviation. Linda answered correctly that the symmetric, bell-shaped graph would have a larger standard deviation. Her reasoning was that "there are more bars. There, um, even though like I think they are clumped pretty well, there's still, um, a higher frequency and so, there's three more, so because of like the extra three there's going to be more like room for deviation just because there's more added values." Other students noted that there were more bars or values in the graph on the right, or that it had a larger range, but still selected the smaller option as an answer. Students tended to note the bell-shape or the perception that more values were in the middle of the right-hand distribution compared to the one on the left as reasons for their responses.

After students checked their answers to item 8 and found out they were incorrect, the interviewer attempted to guide their attention to the ambiguity that stemmed from a comparison of characteristics between the two graphs. Attention was drawn to the differences in the number of values and the range, pointing out that while the bell shape suggested the right-hand graph would have a smaller standard deviation, the larger range made it possible for it to have a larger standard deviation, and the different number of values (or bars) presented a different situation than was presented either during the exploration phase or in the previous tests. The interviewer suggested that ambiguous situations might require calculation of the standard deviations, and demonstrated how the information provided by the program could be used to do so.

When students came to test item 10, they were more likely to note the difference in the number of bars or values between the two graphs and resort to calculating the standard deviation for the graph on the right. Nine of the students came to a correct decision predominantly through calculation. Two of the students (Troy and Jane) initially predicted the graph on the right to have a larger standard deviation, but then noted the discrepancy in range or number of values, calculated the standard deviation, and changed their decisions. Three students (Jeff, Lora, and Linda) did not perform calculations and incorrectly responded that the right-hand graph would have a larger standard deviation. Because Linda correctly answered test item 8, she did not receive the same guidance from the interviewer as the other students, which may account for her incorrect response on test item 10.

4. DISCUSSION

While some research on students' understanding of statistical variability has been conducted, little is known about students' understanding of measures of variability, such as the standard deviation, how that understanding develops, and effective ways to support that development. The study reported here was an attempt to address this lack of information. An exploratory research design was used that employed a multi-phase interview protocol where students interacted with a computer environment designed to help them coordinate underlying concepts that are foundational to an understanding of the standard deviation. The primary instructional goal was to help students develop an understanding of how deviation from the mean and frequency combined to determine the value of the standard deviation. The primary research goal was to develop a better understanding of students thinking about the standard deviation, both after classroom instruction on the concept, and as the concept was developed and elaborated. What has the study contributed with respect to both of these goals? To address this question, a brief critique of some aspects of the research design is offered, followed by a discussion of implications for instruction and future research.

4.1. LIMITATIONS

Definitive statements about the necessity of the instructional design features cannot be made due to the exploratory nature of the investigation. Other approaches may be more optimal and efficient at promoting students' conceptual understanding of the standard deviation. For example, the series of games was designed so that the complexity of the bar arrangements increased incrementally. This was expected to facilitate attending to the various factors that affect the standard deviation and coordination of their simultaneous contributions. However, the incremental increase may not be necessary. The use of several four bar and five bar games may be just as effective in eliciting students justifications and, at the same time, may provide sufficient feedback over a series of games to promote an integrated understanding. It may also be just as effective to describe these relationships to students, perhaps introducing one relationship among factors at a time, building up the complexity of the description in steps, and presenting graphical examples as illustrations for each step.

Students' justifications for their responses to the test items were based on comparisons of relevant characteristics of the histograms and were similar to their justifications expressed during the exploration phase. However, it cannot be firmly established that these justifications were learned or developed during the interview. An independent assessment of the characteristics of histograms students attend to and the type of justifications they give prior to interacting with the computer environment would have been helpful in revealing students' prior knowledge and conceptions.

The situation presented in the computer environment was artificial, although none of the students indicated that this was a problem or produced any difficulty. Nonetheless, a more realistic situation might provide the same type of feedback and support conceptual development while promoting reasoning with real data. The process of building up a distribution, one data value at a time, may be more effective at promoting distributional reasoning and thinking (see Bakker, 2004). For example, values from a real data set could be presented one at a time, and the student could be asked to anticipate the effect on the mean, deviations, and standard deviation of adding the next value to the distribution.

4.2. IMPLICATIONS FOR RESEARCH AND INSTRUCTION

The ensemble of justifications, strategies, and concepts found in this study indicate that students in an introductory statistics course form a variety of ideas as they are first learning about the standard deviation. Some of these ideas, such as the Contiguous, Range, Mean in the Middle, and Far Away-Values rules, capture some relevant aspects of variation and the standard deviation, but may represent a cursory and fragmented level of understanding. Others such as the Far Away-Mean, Balance, More Values in the Middle, and Bell-Shaped rules, represent much closer approximations to an integrated understanding. There are still other ideas, notably the prevalent Equally Spread Out rule and the idiosyncratic Big Mean rule, that are inconsistent with a coherent conception of the standard deviation. Some students also demonstrated an ability to coordinate the effects of several operations on the value of the standard deviation, an indication of a more integrated conception.

To what extent did students achieve a fully coordinated conceptualization of the standard deviation through interaction with the computer program and the interviewer? Only one student, Troy, appeared to have a prior belief that the location of the mean was directly related to the size of the standard deviation, but he readily used mirror images and moving distributions to new locations to produce arrangements with the same standard deviation by the third game. Interaction with the computer environment also appeared effective in changing the conception of students who presented the Equally Spread Out justification early in the interview. By the end of the exploration phase, all of the students appeared to have understood that the mirror image of a distribution conserved the value of the standard deviation. They also demonstrated the understanding that the relative, and not absolute, location of the bars determined the standard deviation.

Only a few students provided justifications by the end of the exploration phase that approached a fully coordinated conception of how frequency and deviation from the mean combine to influence the value of the standard deviation. Students' arrangements and justifications initially indicated an understanding that a large number of values needed to cluster around the mean to produce a relatively small standard deviation, while larger numbers of values needed to be placed far from the mean, in both directions, to produce a relatively large standard deviation. Students' justifications became more complex as the number of bars increased. More students made references to the distribution of values relative to the mean, indicating an awareness of deviation. They also tended to combine earlier justifications, such as presenting both Balance and Far Away arguments to justify a distribution for the largest standard deviation, or combining Contiguity and More in the Middle arguments for why a distribution should have the smallest standard deviation. Therefore, while many students did not have a fully coordinated understanding by the end of the exploration phase, most had developed parts of this conceptualization and began to coordinate some of the concepts.

Most of the students used a rule-based approach to compare variability across distributions instead of reasoning from a conceptual representation of the standard deviation. Even among the students with apparently richer representations, their explanations during the testing phase were usually based on finding a *single* distinguishing characteristic between the two distributions rather than reasoning about the size of the standard deviation through a conception that reflected how density was distributed around the mean in each distribution. This suggests that students tend to take a rule-based, pattern recognition approach when comparing distributions. If this is the case, two questions need to be addressed. What type of experiences are required to move students from a rule-based approach to a more integrated understanding that can be generalized to a variety of contexts? In addition to the interactive experience presented in the current study, do students need additional support to reflect on the relationships between the different factors and to attend to and coordinate the related changes?

A pattern-recognition, rule-based approach is consistent with a goal of finding the right answer by noting characteristics that differentiate one arrangement from another and noting the correspondence with the size of the standard deviation. This orientation was supported by the design of the software in that the value of the standard deviation was always available and there were no consequences for an arrangement that did not meet a criterion. This may have overemphasized the need to be "correct" instead of promoting the exploration and reflection needed to develop an understanding of factors that affect the standard deviation. Is there a way to modify the task so there is less emphasis on a correct solution and more emphasis on exploring the relationships among the factors that affect the standard deviation? One possibility is to modify the software so that the mean and standard deviation are not automatically revealed. This might promote reflection and result in less Guess and Check behavior.

The interviewer attempted to extend students' conceptual understanding by trying to draw their attention to relevant aspects of the distributions, and by modeling the desired conceptual understanding. The software was designed to help students identify factors that affect the standard deviation. The software and interview were not designed with the promotion of model building in mind. A model eliciting approach may be more likely to produce the "system-as-a-whole" thinking (Lesh & Carmona, 2003) that is needed for a fully coordinated conception of the standard deviation, and to allow students to develop a more integrated representational system (Lehrer & Schauble, 2003), rather than a collection of separate and potentially conflicting rules.

Several changes to the program could be introduced to support model building and study how it affects understanding of the standard deviation. The software currently draws attention to a single bar rather than visually emphasizing how characteristics change simultaneously. A second display above the graphing area of the histogram that presents horizontal deviation bars colored to match the corresponding vertical frequency bar colors may facilitate coordination of simultaneous changes in values, the mean, deviations, and the standard deviation. The interview protocol would also need modification to include model eliciting prompts and probes. This can be done through eliciting conjectures from students about how changes to arrangements will affect the value of the mean and deviations, and how these changes subsequently affect the standard deviation, promoting a coordination of the concepts and a relational structure that models their mutual effects. This contrasts with the current protocol where the interviewer modeled the thinking and reasoning for the student rather than supporting students to produce their own conjectures and test their implications.

ACKNOWLEDGMENTS

The authors would like to thank three anonymous reviewers and the editor for their careful and thoughtful reviews of earlier versions of this manuscript.

REFERENCES

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools. (Doctoral dissertation, Utrecht University). Utrecht: CD-β Press.

Bubules, N. C., & Linn, M. C. (1988). Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology, 80*, 67-75.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chin, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49.

delMas, R. (1997). A framework for the development of software for teaching statistical concepts. In J. B. Garfield & G. Burril (Eds.), *Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 International Association of Statistics Education (IASE) Round Table Conference* (pp. 85-99). Voorburg, The Netherlands: International Statistical Institute.

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

delMas, R., Garfield, J., & Chance, B. (2004). Using assessment to study the development of students' reasoning about sampling distributions. Paper presented at the Annual Meeting of the American Educational Research Association, April 13, San Diego, CA. [Online: www.gen.umn.edu/faculty_staff/delmas/AERA_2004_samp_dist.pdf]

Garfield, J., delMas, R., & Chance, B. (1999). The role of assessment in research on teaching and learning statistics. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Hardiman, P., Pollatsek, A., & Well, A. D. (1986). Learning to understand the balance beam. *Cognition and Instruction, 3*, 63-86.

Hoerl, R., & Snee, R. D. (2001). *Statistical thinking: Improving business performance*. Duxbury Press: Pacific Grove, CA.

Lehrer, R., & Schauble, L. (2003). Origins and evolution of model-based reasoning in mathematics and science. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 59-70). Mahwah, NJ: Lawrence Erlbaum.

Lesh, R., & Carmona, G. (2003). Piagetian conceptual systems and models for mathematizing everyday experiences. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 71-96). Mahwah, NJ: Lawrence Erlbaum.

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

Moore, D.S. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants*. Washington, DC: National Academy Press.

Nickerson, R. S. (1995). Can technology help teach for understanding? In D. N. Perkins, J. L. Schwartz, M. M. West, & M. S. Wiske (Eds.), *Software goes to school: Teaching for understanding with new technologies* (pp. 7-22). New York: Oxford University Press.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66(2)*, 211-227.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Roth, K., & Anderson, C. (1988). Promoting conceptual change learning from science textbooks. In P. Ramsden (Ed.), *Improving learning: New perspectives* (pp. 109-141). London: Kogan Page.

Saldahna, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270.

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Bidulph & K. Carr (Eds.), *Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 6-22). Rotorua, N.Z.: University of Waikata.

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

Snee, R. (1990). Statistical thinking and its contribution to quality. *The American Statistician, 44*, 116-121.

Snir, J., Smith, C., & Grosslight, L. (1995). Conceptually enhanced simulations: A computer tool for science teaching. In D. N. Perkins, J. L. Schwartz, M. M. West, & M. S. Wiske (Eds.), *Software goes to school: Teaching for understanding with new technologies* (pp. 106-129). New York: Oxford University Press.

Thompson, P. W., Saldahna, L. A., & Liu, Y. (2004). Why statistical inference is hard to understand. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

ROBERT DELMAS
University of Minnesota
354 Appleby Hall
128 Pleasant Street SE
Minneapolis, MN 55455
USA

THINKING TOOLS AND VARIATION

MAXINE PFANNKUCH
The University of Auckland, New Zealand
m.pfannkuch@auckland.ac.nz

Editor's note: This article is a discussion of and reaction to five papers which appeared in the SERJ special issue 3(2) published in November 2004, on Research on Reasoning about Variation and Variability. These include four refereed papers reporting original research, by Hammerman and Rubin, Ben-Zvi, Bakker, and Reading, and Gould's invited paper providing a statistician's view of variation and its analysis. For brevity, these papers will be referenced here only by author names, without the year of publication.

SUMMARY

This article discusses five papers focused on "Research on Reasoning about Variation and Variability", by Hammerman and Rubin, Ben-Zvi, Bakker, Reading, and Gould, which appeared in a special issue of the Statistics Education Research Journal (No. 3(2) November 2004). Three issues emerged from these papers. First, there is a link between the types of tools that students use and the type of reasoning about variation that is observed. Second, students' reasoning about variation is interconnected to all parts of the statistical investigation cycle. Third, learning to reason about variation with tools and to understand phenomena are two elements that should be reflected in teaching. The discussion points to the need to expand instruction to include both exploratory data analysis and classical inference approaches and points to directions for future research.

Keywords: Statistics education; Variation; Reasoning; Thinking tools; Statistical investigation; Exploratory data analysis

1. INTRODUCTION

The five papers in the SERJ November 2004 special issue on reasoning about variation and variability not only open a window into an exciting new research area but also give an excellent insight into current research on developing cognitive and pedagogical theories. The opportunity to read how statisticians, teachers, and students reason about variation proved to be interesting and thought–provoking. Each of the papers gives its own unique perspective into and interpretation of reasoning about variation, whereas when considering the five papers as a whole the research on variation reasoning becomes more than the sum of all these findings. It then becomes a challenge to explicate and to extract some themes and issues about variation from all the papers.

Variation lies at the heart of statistical enquiry and is one of the "big ideas" that should underpin statistical teaching (Moore, 1990; Gal & Garfield, 1997). Consideration of the effects of variation influences all thinking and reasoning through every stage of the empirical enquiry cycle – problem, plan, data, analysis, and conclusion (Pfannkuch & Wild, 2004). Thinking about variation starts from noticing variation in a real situation, the problem. It influences the strategies that are adopted in the planning and data collection stages, for example by introducing planned variation into data production through the use of randomization or reducing known sources of variability. In the analysis and

Statistics Education Research Journal, 4(1), 83-91, http://www.stat.auckland.ac.nz/serj

conclusion stages the presence of variation determines how perceived patterns might be characterized and acted upon.

Research on how students might develop skills to reason about variation and variability throughout the empirical enquiry cycle has been limited (Shaughnessy, 1997). Further, although variation is a core concept in statistics, teaching has not focused on enculturating students into variation-type thinking and reasoning approaches. Much needs to be learnt about how to cultivate and develop this type of reasoning in students. The November 2004 of SERJ will be seen as a landmark special issue in establishing some beginning foundations for such a knowledge base, and is followed by additional papers on the same topic in the current issue.

With the juxtaposition of a statistician's (Gould's) view of how he deals with and reasons about variation, and the views of statistics education researchers of how students and teachers reason about variation (Bakker; Ben-Zvi; Hammerman & Rubin; Reading), many interesting avenues for enquiry are suggested. One noticeable feature of the five papers is the different thinking tools used by the statistician and students and teachers for the analysis of the data, such as graphs, tables of data, or diagrams. I view these as "thinking tools" because they are more than mere external representations; for example, in the analysis stage they allow the statistician and students to interrogate the data, to extract information contained in the data, and to think and reason with the data. Another noticeable feature is that the focus is on the variation in the *analysis* stage of the empirical enquiry cycle. Ideally teaching and learning data analysis should always be situated within the full empirical enquiry cycle. While specific grade levels may focus on developing particular statistical concepts, the overall cycle of enquiry should always be part of the teaching approach, otherwise students' thinking and reasoning can become disconnected (Pfannkuch & Wild, 2003). These two noticeable features raise two main issues. First, about the relationship between tools and the types of reasoning that they encourage or permit and second, about how reasoning about variation is interconnected to all parts of the enquiry cycle.

What strikes me most, however, from the juxtaposition of the statistician paper and education research papers, is the contrast in the purposes of investigating data. The nature of the analyses that the students are doing seems to me to be leading towards the future enculturation of students into classical inference. In comparison, the statistician's focus, in the statistical case studies described, is more about modeling variation to arrive at insights about underlying data-generated mechanisms. The thinking is less structured, more wide-ranging, and draws on many knowledge sources. Gould's focus is on modeling variation for the purposes of explanation (Lead-Level case study) and prediction (UCLA Rain, Drinking Trends and Chipmunk case studies). This Exploratory Data Analysis (EDA) approach (cf., Tukey, 1977; Shaughnessy, Garfield, & Greer, 1996), whose main aim is modeling and understanding phenomena, is in stark contrast to the simple estimation of parameters of populations from samples, which underscores most of current statistics education.

In light of the above, three main issues are addressed in this paper: the thinking tools used for displaying variation for analysis and the type of reasoning observed; the need to consider how reasoning about variation involves the whole empirical enquiry cycle; and the need to extend the statistics curriculum to include analytic or non-confirmatory situations. Questions will be raised in response to the papers and possible directions for future research will be explored.

2. THINKING TOOLS IN THE ANALYSIS STAGE

The five papers use a range of thinking tools in the analysis stage of the empirical enquiry cycle, such as tabular graphs, graphs, tables of data, and diagrammatic representations, to aid students' reasoning about variation. This section will examine what each paper presents in this regard, reflect on the findings about students' reasoning, and raise questions about the first main issue – possible interconnections between the tools used and the reasoning observed.

Hammerman and Rubin describe how teachers used a new dynamic data visualization tool (Tinkerplots$^{\text{TM}}$) to divide distributions into slices and consequently compared frequencies and percentages within these slices to make inferences. This tool, which I will call *the tabular graph* approach, is an integration of table and graph features. The type of thinking observed was slice-wise comparison across groups, which tended to ignore the distribution as a whole (see also Cobb, McClain, & Gravemeijer, 2003). According to Hammerman and Rubin, teachers were viewing the data from a classifier perspective, through their focus on comparing frequencies of particular attribute values within and between data sets, rather than seeing the data sets as a whole, a pre-cursor stage for aggregate-based reasoning. They argue that this new tool engendered and made visible thinking that had previously lain dormant or invisible, since the teachers' slice-wise comparison reasoning seemed to be an extension of pair-wise comparison reasoning that other researchers have documented (e.g., Moritz, 2000), whereby students compare two individual cases. The tool provided the means to divide the data up, but reasoning about variability also requires detecting patterns in the whole distribution. Is this new integrated tabular graph tool helping or hindering understanding? If the visible frequency count bars or unorganized dots in the bins were concealed by a bar graph in this tool, would the teachers' thinking change? Did this tool produce cognitive overload in the teachers with one of the consequences being subgroup or bin reasoning? Or has this tool opened up completely new ways of reasoning about variation?

Ben-Zvi gives a thorough insight into how two students grappled with variability within and between two distributions. The students built up their inferential reasoning by being introduced to a sequence of tools throughout the learning process, namely frequency table, percentage table, statistical measures, graph. This is a *tables plus graph* approach. Ben-Zvi's approach is to let the students focus and think with one tool at a time. First, with the table, the students were scaffolded to notice the variation and then formulate a hypothesis. To justify their hypothesis they argued with a table of percentages where slice-wise comparisons were made, as well as range comparisons involving variation within and between two sets of data in table format. The introduction of statistical measures seemed to produce few new steps in their thinking. The production of a series comparison graph by the students induced pairwise comparison of the bars, which led to seeing the pattern in the variation and to articulating a trend. Ben-Zvi gets his students to think with the table but other research (e.g., Pfannkuch & Rubick, 2002) shows students do not necessarily do this since the table is perceived as a means to organize data for graphing, not as a tool to think with. Did using the table aid the students' thinking? Was the percentage table a necessary precursor to their thinking with graphs? Did these students integrate the percentage table and graph into their thinking or did the thinking remain separated? How should the teacher integrate statistical measures into their graphical reasoning? Should thinking tools be introduced in a sequential way?

Bakker, who deals with sample distributions and "grows" them to population distributions, has a purely *graphical or diagrammatic reasoning* approach. The students described the variability in predicted sample data distributions, comparing their predictions to real sample data sets. The focus is on transitioning students' informal language and notions to statistical language and notions and on transitioning the students from seeing variability in data plots to seeing variability in continuous distributions. The aim is to stay away from the data and foster the notion and concept of distribution by discussing averages and spread in relation to shape. An important part of using these graphical tools for reasoning is the role of reflection and the role of the teacher in facilitating students' conceptions. However, Bakker is not sure whether students recognized that there was more variability in small samples. What patterns were students conceiving within the variability of small samples? How did they integrate into their reasoning the sample distributions with the population distribution? If the students were given a population distribution and asked to take a sample of size 10 from it what would their predicted sample distribution look like?

Reading reports that her students did not use graphic tools for reasoning but rather used *tables of data*. The students noticed the variation, and looked for and detected patterns in the data. This is a natural human response, since all humans tend to see patterns even if none exist (Wild & Pfannkuch, 1999) and to find causal connections (e.g., Watson, Collis, Callingham, & Moritz, 1995). Reading distinguishes a hierarchy of reasoning from qualitative to quantitative. In the qualitative responses the students noticed and acknowledged the variation, whereas at the quantitative level there were emergent ideas of measuring and quantifying the variation such as reasoning with the minima, maxima, ranges, and notions of deviation. Yet, the teacher did not scaffold Reading's students' reasoning and few students chose to use graphs (with which they were familiar). What were the characteristics of the data table that induced students not to change the data to another representation?

In contrast, Gould, the statistician, uses multiple graphs and models as his thinking tools and keeps re-organizing the data to draw out the information contained in the variation. This is the transnumerative type of thinking promoted by Wild and Pfannkuch (1999). He chips away at the unexplained variation to gain more information about patterns in the data and about underlying factors that may be sources of variation in the realistic situations examined. He has a data detective approach – searching for patterns and trends, causes, and explanations for the variation.

These studies suggest that there is a link between the thinking tools used and the reasoning observed in the students or teachers studied. This raises the question as to what tools or sequence of tools will aid students or teachers to conceive both statistical patterns and deviations from those patterns. Prompting students to use multiple models or graphs, multiple ways of seeing and articulating patterns, would seem to be one solution suggested by this research. We do not know, however, how students or teachers link and interconnect their reasoning across tools or how the tools affect, limit, or empower their thinking.

With the above in mind, it is useful to look more broadly at the notion of a thinking tool. In mathematics education research, there is wide recognition that thinking tools or representations do affect students' thinking. Mesquita (1998) observes how the role and nature of the external representation in geometry affects students' thinking. She identifies external representations as having two roles, descriptive and heuristical. The descriptive representation "illustrates multiple relationships and properties involved in the problem without suggesting solution procedures" (p. 191) whereas the heuristical representation "acts as a support for intuition, suggesting transformations that lead to a solution" (p. 191). In another study, Hegedus and Kaput (2004) argue that their dynamic connective representations for algebra activities accounted for significant gains in student performance, which provides further evidence of a link between students' thinking and the thinking tool employed. Mesquita believes that external representations are relevant elements in cognitive theory. In fact, Thomas (2004) reports that there is no coherent cognitive theory on the role and nature of external representations.

Fischbein (1987, p. 125) states that thinking tools or models are useful for the learner to gain access to an understanding of a concept. For productive reasoning the tool must "constitute an intervening device between the intellectually inaccessible and the intellectually acceptable and manipulable." He warns, however, that the properties inherent in the tool may lead to an imperfect mediator and hence can cause incomplete or incorrect interpretations. In the light of what Fischbein is saying the results from the empirical investigations reviewed here suggest that some statistical thinking tools developed by statisticians may be considered as imperfect thinking mediums and may lead to misinterpretation. One avenue to overcome such misinterpretations is for the teacher and learner to become aware of these intuitive obstacles. Another avenue to pursue is that advocated by Fischbein (1987, p. 191): "to create fundamentally new representations." Therefore statistics educators and statisticians may need to reassess the existing thinking tools in the discipline and perhaps create new thinking tools that are more closely aligned to human reasoning, an approach taken by the Tinkerplots™ software creators Konold and Miller (2004).

3. VARIATION AND THE EMPIRICAL ENQUIRY CYCLE

Statistical investigations are conducted to expand contextual knowledge and to improve understanding of situations, as well as to solve data-based problems. In reporting on statistical investigations, usually the problem and the measures taken are defined, the method of data collection is explained, and the main results are communicated. It is important, however, to ask how much information is provided to learners about the *context* from which data emerged, and what role does contextual information have. Gould, the statistician, briefly gives details of where his four sets of data have come from, how data were collected and how measurements were taken. For the reader this allows extra thinking tools for understanding his data and findings. This raises the issue as to whether lack of such information might inhibit or prevent students' thinking from reaching an optimal level. This type of information feeds into the analysis and conclusions parts of the cycle. In order to demonstrate how seeking explanations for the variation is grounded in the empirical enquiry cycle, questions will be raised below about the sources of the data for each of the statistics education papers. These remarks do not assume that these aspects were neglected in the four classroom studies; rather, they aim to raise the second main issue about considering how students' reasoning about variation is interconnected with the whole enquiry cycle.

Ben-Zvi analysed reactions to a task which asked students to compare the lengths of Israeli and American last names. Information that might help an analysis would be: Where did these names come from? What parts of the respective countries? What grades were the two classes selected? What year were they collected? Bakker states that the teacher used real data sets about the weights of groups of children. Were they eighth graders from another Dutch school? Did the school have the same socio-economic level, same ethnicity as the students' school? Were the weights of one gender or both genders? How accurate were the data collected—self-report or actual weighting with a scale?

Hammerman and Rubin ask their teachers to compare homework hours studied by students at two schools. Before comparing the data it would be helpful to know the following: Were the students the same age? How were the data collected – recall of number of hours for one particular week, diary over one week, or was it students' perception of how much homework they usually did each week? Were the two schools similar in ethnicity, socio-economic level? Were both genders included? Was a random sample taken from each school or were particular classes selected? Reading has students analyse real weather data from the students' town. She has a three-year census of data but each student is given a sample of one month's rainfall and later on – in another teaching segment – one month's temperature data to comment on. Questions that need to be considered are: How were these data measured? Were they averages from a number of weather stations or from one weather station? In what parts of the town were these measurements taken? How is a wet day defined in that area? What is the difference between a "wet" day and a "showery" day and how is this information captured in the rainfall measurement? What range of temperatures is considered best for outdoor activities?

The variation that is considered and dealt with in the first two stages of the empirical cycle, problem and plan, should be part of the information or story that is presented to students. One could argue that there would be too much information for students to grasp, hence there should be a balance between providing enough background information to inform the analysis and enable students to engage with their data, versus providing no information, which may leave teachers wondering why students did not consider certain questions. Hammerman and Rubin notice that the analyses appeared to be context dependent. Their teachers seemed to reach an aggregate perspective with the AIDS data and one wonders whether the teachers were given or had more background information about the data.

All the above issues about the problem and plan stages may well have been discussed with the classes but the statistician included his story in his paper while the researchers left their stories out and the question is why? Is it just that their focus was different when writing their papers? Or is teaching different from the practitioner way of doing data analysis? Or should teaching incorporate more of the statisticians' practice into their teaching? Whatever the answer, research that explores students' reasoning about variation with and without the investigative cycle information may provide insights into pedagogical practice and students' cognition.

A secondary but related issue to the above discussion is that the thinking tools used by statisticians and hence statistics educators tend to be tools for the analysis stage of the empirical enquiry cycle. Wild and Pfannkuch (1999), however, argue that new thinking tools need to be created and used for the other stages of the cycle. The quality management field has created thinking tools for the enquiry cycle such as Joiner's (1994) seven-step process for projects, which gives lists of critical questions and diagrams to prompt, stimulate, and trigger thinking. Quality management defines statistics in terms of variation, and hence variation underpins all its thinking tools. Such tools are successfully used in the quality management field and there seems to be no reason why the statistics discipline and statistics education researchers should not adapt and expand these tools for the general field (see Section 4 of Wild & Pfannkuch, 1999).

4. PURPOSE OF THE STATISTICS CURRICULUM

The third main issue raised upon reading the five papers is that the nature of the analyses seems to be leading students towards a classical inference perspective. This raises the question about the purposes of the thinking tools that were used by the students. EDA and classical inference should be two major strands of statistical investigation in a modern curriculum. It was W. E. Deming in 1953 who first raised the distinction between what he termed enumerative and analytic studies. Enumerative studies are concerned with describing the current situation and inference is limited to the population or process sampled. Current statistics education curricula seem to me to be geared towards formal statistical inference and the carrying out of tests to determine information about the population from the sample data. Causal thinking or seeking explanations or making predictions for phenomena are inadvertently discouraged in such an approach, as probabilistic thinking is a key outcome. Analytic studies involve using data from current processes or populations for understanding and seeking explanations for observed phenomena in order to control or predict the future behavior of processes or populations. It is the measuring and modeling of variation for these purposes as well as explaining and dealing with it that Wild and Pfannkuch (1999) considered important. Gould, the statistician, demonstrates such an approach in his case studies as he attempts to learn more about real world situations through interacting and interrogating the data and to provide some possible solutions (e.g., the identification of chipmunk families).

The differentiation between analytic and enumerative studies suggests a dual approach to reasoning about variation. Consideration of variation should allow for the seeking of explanations, looking for causes, making predictions, or improving processes. Yet at the same time awareness that not all variation can be explained should be raised. That is, there is a need to determine whether the observed differences in sample data reflect the underlying differences in populations, the explained systematic variation, or are due to sampling variability—the unexplained variation statisticians choose to model as random variation. An analytic approach involves learning more about observed phenomena. Each statistics education paper is now considered to raise questions about what the students might have learnt when they interacted with their data if they were afforded an EDA perspective. Again, these remarks do not assume that the EDA perspective was not adopted. Rather, the issue raised is the need to query how teaching could adopt both EDA and classical inference approaches.

In Ben-Zvi's students' conceptions of variability in data, they used their contextual knowledge to realize that a source of variation was in the structure of the two languages. This interaction between data and context enabled them to continue to believe that there was a difference between the two sets of data. What did his students learn about the real world situation? Perhaps they became aware that English surnames are usually longer than Israeli surnames a fact that they had not considered before. Ben-Zvi mentioned that the sources of variation were cultural, historical, and hereditary but he did not seem to explore these ideas with his students.

Bakker's students, when reasoning about the shape of a weight distribution, used their contextual knowledge of weight to explain the type of variability they would expect. Bakker's students were not using actual data when discussing a graph shape for weight; rather, they used "everyday data" that are collected as one operates in the daily environment. They were interrogating data that had not been collected in a planned way. Tversky and Kahneman (1982), Snee (1999), and Pfannkuch and Wild (2004) believed it was possible to have a statistical perspective and understand variation when not being able to collect data. What these students were learning was how to reason about variation using their contextual knowledge. Contextually they learned that weight has a positively skewed distribution, that is, for the Dutch eighth graders population, weights which are much greater than average are more common than weights which are much smaller than average.

Reading facilitates her students to make predictions about future behavior, by requiring students to look at one month of weather data and to argue whether that particular month would or would not be a good time to hold a youth festival. Her students were at the stage of noticing variability within the data and arguing verbally with scant summarization and modeling of data. It seemed that the students were relying on their experience of the context or their "everyday" data that they had collected on the weather and not the information in the data. This raises another question as to how do students learn when to argue by referring to or using the given data, their "everyday" data, or both? What did her students learn about the real world situation? Presumably they learnt about variability present in the weather during one month. Sources of variation such as the season, year, and data collection method could be considered for further exploration of the data as well as wondering if they took the same month of data in different years whether they would detect similar patterns in the data. From the classical inference perspective the students could consider that they had a sample from a population. Hence, with these weather data both an EDA and classical inference approach could be adopted.

Hammerman and Rubin describe teachers comparing slices when making decisions about whether there was a difference between groups. The teachers reduced the variability in the data by continually breaking up the data into categories. In particular, when comparing hours of homework studied between two groups, the teachers argued with a subgroup of the data to state that Amherst students studied more homework than the Holyoke students. If these hours of homework studied were considered, then hypothetically the next step would be to seek an explanation for the variation between the two schools. If there was a belief or research evidence that hours of homework were linked to improved performance, a specific improvement in the Holyoke system, based on the explanation of the variation, could be implemented in order to raise the level of homework hours per week. Another question about homework hours was to consider whether the difference was real or whether it was due to sampling variation. Herein lays the duality of classical inference and EDA. Assuming that the number of homework hours was real data and that the teachers knew the schools and students since they explained the upper end variation by commenting on "lifestyle", what did these teachers learn about the real situation? When they looked at the "typical student" and found Amherst students did more homework there seemed to be no explanation for this source of variation. It was not clear whether this finding was a revelation to the teachers or a confirmation of what they already knew.

5. CONCLUSIONS AND FURTHER RESEARCH

Variation is at the core of statistical thinking (Moore, 1990) and the research papers reviewed are beginning to uncover how variation reasoning might be integrated into and revealed within the statistics discipline. The task for the students and teachers who participated in the studies reviewed was noticing and distinguishing variability within and between distributions in many forms, such as:

- tabular graphs which can be an integration of information represented as frequencies, percentages, unorganized dots, or count bars;
- graphs which can be dot plots, continuous distributions, or series comparison graphs;
- tables which can contain raw data, frequencies, or percentages.

The question remains about how to enculturate students into reasoning about variation with a variety of analysis thinking tools, as well as into focusing on learning more about observed phenomena. Perhaps Gould provides an avenue for consideration when he states that confirmatory analysis should be de-emphasized and more attention paid to the noise.

Overall, the five papers examined here will contribute greatly towards variation reasoning becoming an integral part of statistics students' experience and thought. The juxtaposition of these papers raised three key issues which have been the focus of this discussion.

The first key issue is the link between the types of tools students were exposed to and the type of reasoning about variation that was observed. Future research should explore how representations affect, limit, or empower students' thinking; particularly as statistics learning may require representations that suggest multiple ways of transforming and reorganizing data in order to seek out the information contained in the data. For example, the integrated tabular graph representation does not exist as a conventional statistics tool. Therefore research could explore whether the tools actually assist in enculturating students towards reasoning with conventional statistical thinking tools. A comparison of students' reasoning about variability between the Hammerman and Rubin integrated approach and the Ben-Zvi sequential approach is another possible research avenue. Furthermore, the integrated tabular graph representation that was explored by Hammerman and Rubin has not featured in mathematics education research (Thomas, 2004) and therefore statistics education research should be able to make a contribution to cognitive theories on external representations.

The second key issue is the grounding of students' reasoning about variation within the whole enquiry cycle. Research should not only explore students' reasoning about variation and variability within and between each stage but also consider how the reasoning is connected to the overall cycle. Researchers could also develop thinking tools for all parts of the empirical enquiry cycle and then explore how students reason about variation using these tools.

The third key issue is the need to extend teaching approaches and the curriculum towards an EDA perspective. Herein a tension arises in two goals for learning: learning to use a tool to reason about variation; and reasoning about variation to understand a phenomenon. Thinking of sources of variability and then re-organizing the data to produce more distributions to reason with is part of the process. Reasoning about variation involves the interrogation of the data and requires the interplay between context and the variation in data. The context can be general knowledge about the subject area, knowledge of how the data were collected, how the measures were defined, or the "everyday" data that one collects. Reasoning about variation also involves detecting patterns through the interplay between the centre or trend and the variation. All the data interrogated by these students could have led to them understanding and learning more about an observed phenomenon. Unlocking the story in the data requires both an analytic EDA and an enumerative classical inference approach. Future research needs to consider how students can be enculturated into both ways of reasoning and in particular examine how students reason about variation for the purposes of explanation and prediction.

REFERENCES

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*(1), 1–78.

Fischbein, E. (1987). *Intuition in Science and Mathematics*. Dordrecht, The Netherlands: Reidel.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1–13). Amsterdam: IOS Press.

Hegedus, S., & Kaput, J. (2004). An introduction to the profound potential of connected algebra activities: Issues of representation, engagement and pedagogy. In M. Hoines & A. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 129-136). Bergen, Norway: Bergen University College.

Konold, C., & Miller, C. (2004). *Tinkerplots™ Dynamic Data Explorations*. Emeryville, CA: Key Curriculum Press.

Joiner, B. (1994). *Fourth generation management*. New York: McGraw-Hill Inc.

Mesquita, A. (1998). On conceptual obstacles linked with external representation in geometry. *Journal of Mathematical Behavior*, *17*(2), 183-195.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, D.C.: National Academy Press.

Moritz, J. B. (2000). Graphical representations of statistical associations by upper primary students. In J. Bana & A. Chapman (Eds.), *Mathematics education beyond 2000: Proceedings of the Twenty-third Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 440-447). Sydney: MERGA.

Pfannkuch, M., & Rubick, A. (2002). An exploration of students' statistical thinking with given data. *Statistics Education Research Journal*, *1*(2), 4–21. [Online: www.stat.auckland.ac.nz/serj]

Pfannkuch, M., & Wild, C. J. (2003). Statistical thinking: How can we develop it? In *Bulletin of the International Statistical Institute 54th Session Proceedings, Berlin, 2003* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

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

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of chance and data. In F. Biddulph & K. Carr (Eds.), *People in mathematics education: Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 6-22). Sydney: MERGA.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (Part1, pp. 205–237). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Snee, R. (1999.) Discussion: development and use of statistical thinking; a new era. *International Statistical Review*, *67*(3), 255–258.

Thomas, M. (2004). Personal communication, 14 October 2004, Auckland, New Zealand.

Tukey, J. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tversky, A., & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3–20). New York: Press Syndicate of the University of Cambridge. (Originally published in *Science*, *185*(1974), 1124–1131)

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

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, *67*(3), 223-265.

MAXINE PFANNKUCH
Department of Statistics
The University of Auckland
Private Bag 92019
Auckland
New Zealand

A FRAMEWORK FOR TEACHING AND ASSESSING REASONING ABOUT VARIABILITY

JOAN GARFIELD
University of Minnesota
jbg@umn.edu

DANI BEN-ZVI
University of Haifa
dbenzvi@univ.haifa.ac.il

SUMMARY

This article is a discussion of and reaction to two collections of papers on research on Reasoning about Variation: Five papers appeared in November 2004 in a Special Issue 3(2) of the Statistics Education Research Journal (by Hammerman and Rubin, Ben-Zvi, Bakker, Reading, and Gould), and three papers appear in a Special Section on the same topic in the present issue (by Makar and Confrey, delMas and Liu, and Pfannkuch). These papers show that understanding of variability is much more complex and difficult to achieve than prior literature has led us to believe. Based on these papers and other pertinent literature, the present paper, written by the Guest Editors, outlines seven components that are part of a comprehensive epistemological model of the ideas that comprise a deep understanding of variability: Developing intuitive ideas of variability, describing and representing variability, using variability to make comparisons, recognizing variability in special types of distributions, identifying patterns of variability in fitting models, using variability to predict random samples or outcomes, and considering variability as part of statistical thinking. With regard to each component, possible instructional goals as well as types of assessment tasks that can be used in research and teaching contexts are illustrated. The conceptual model presented can inform the design and alignment of teaching and assessment, as well as help in planning research and in organizing results from prior and future research on reasoning about variability.

Keywords: Variability; Variation; Learning and teaching statistics; Assessment; Statistics curriculum; Statistics education; Reasoning and thinking; Deep understanding

1. INTRODUCTION

Variability is at the heart of statistics and is the fundamental component of statistical thinking (Pfannkuch, 1997; Pfannkuch & Wild, 2004; Shaughnessy, 1997). Variability is what makes decisions in the face of uncertainty so difficult. Variability is what makes statistics so challenging and interesting and allows us to interpret, model and make predictions from data (Gould, 2004). Therefore, variability should be repeatedly integrated, revisited and highlighted in statistics curriculum and instruction (e.g., Moore, 1990). Moore (1992, p. 426) extends this notion of the centrality of variation by stating that “pupils in the future will bring away from their schooling a structure of thought that whispers ‘variation matters’.” Variability should be centrally emphasized from the earliest grades (in formal and informal activities and discussions) through high school and the introductory college course. Suggestions for how this should be done and evaluated are the focus of this paper.

Statistics Education Research Journal, 4(1), 92-99, http://www.stat.auckland.ac.nz/serj

This article discusses and extends ideas appearing in two collections of papers on research on Reasoning about Variation. Five papers appeared in the November 2004 special issue of the *Statistics Education Research Journal* (Hammerman & Rubin, Ben-Zvi, Bakker, Reading, Gould), and three papers appear in the Special Section on the same topic in the present issue (Makar & Confrey, delMas & Liu, Pfannkuch). These two collections of papers overall provide us with a strong case that understanding of variability is much more complex and difficult than prior literature suggests. For example, even college students or teachers, examined in some of the studies, demonstrate diverse intuitions, misconceptions, and incomplete or shallow understanding (Hammerman & Rubin, 2004; delMas & Liu, 2005; Makar & Confrey, 2005). Shaughnessy, Watson, Moritz, and Reading (1999) have similarly found a lack of clear growth in students' conceptions of variability for a particular task. A full and complete understanding of variability means developing a cognitive model that includes several components and their connections, and using this model to reason about variability in different contexts.

The two collections of papers present general ideas regarding ways for teaching and assessment related to knowledge of variability, but have not examined in much detail bridges to practice, in the form of specific instructional goals and assessment methods. Hence, building from the foundations of the collected set of issues these papers discussed, as well as from our own work, in this paper we suggest an epistemological model and associated actions and behaviors that would be demonstrated when students reason about variability as they encounter data in solving statistical problems. In the following sections, we offer a preliminary but extensive set of ideas both regarding instructional topics as well as regarding assessment tools associated with the model of reasoning about variability, and close with a discussion of implications of the suggested model for practice and future research.

2. DEVELOPING DEEP UNDERSTANDING OF VARIABILITY

In this section, we outline the components of an epistemological model that consists of a set of ideas loosely grouped into seven overlapping categories or areas. These areas can be viewed as building blocks for constructing "deep understanding" of the complex concept of variability. A deep understanding means that instead of being able to recite only fragmented pieces of information, students develop relatively systematic, integrated or holistic understandings. Mastery is demonstrated by their success in producing new knowledge by discovering relationships, solving new problems, constructing explanations, and drawing conclusions. Deep understanding is robust, reflective, i.e., includes awareness of own boundaries, of limits of applicability, and also operable, i.e., can be used to solve real problems. It is characterized by its breadth, by knowing the concept components and their connections, and most importantly, by knowing what is important and what is less important or irrelevant. Deep understanding is similar to "conceptual understanding" and to Skemp's "relational understanding", as opposed to "procedural understanding" (Skemp, 1976). Conceptual understanding similarly implies knowledge of the idea, and how it relates to already acquired ideas. It also requires an understanding of the contexts within which the idea is applicable, as well as its limitations. It enables a person to apply and adapt an idea flexibly to new situations rather than just following learned procedures in familiar situations.

Below we present key ideas in each of the seven areas of knowledge of variability:

1. Developing intuitive ideas of variability

- Recognizing that variability is everywhere (the omnipresence of variability; Moore, 1990, 1997). Individuals vary on many characteristics, and repeated measurements on the same characteristic are variable. Both qualitative and quantitative variables reveal variability of data.
- Some things vary just a little, some vary a lot.
- We can try to understand why things vary: By thinking about and examining the variables we can try to explain the different reasons and sources for variability.
- Variability is a general or global characteristic of a data set. It involves considering data as an entity, rather than as individual points or as a combination of center and extreme values.

2. Describing and representing variability

- Graphs of data show how things vary and may reveal patterns to help us focus on global features of distributions and identify the signal in the noise.
- Different graphs may reveal different aspects of the variability in a data set so it is important to study more than a single graph of a data set.
- We can use one number to represent a global feature (such as variability) of the distribution.
- Different numerical summaries tell us different things about the spread of a data set. For example, the Range tells us the overall spread from highest to lowest value, while the Standard Deviation (SD) tells us the typical spread from the mean. The Interquartile Range (IQR) tells us the spread of the middle half of a distribution.
- While the IQR and SD tell us about variability of data, they are most useful for interpreting variability when we also know the related measure of center (mean for SD, median for IQR) as well as the general shape of the distribution.
- Measures of variability and center (as long as we consider them together) are more or less informative for different types of distribution. For example, the mean and SD tell us useful information about symmetric distributions, in particular, the normal distribution. For skewed distributions, the median and IQR are more useful summaries.

3. Using variability to make comparisons

- When making comparisons of two or more data sets, examining their graphs on the same scale allows us to compare the variability and speculate about why there are differences in the data sets.
- It is helpful to use global summaries of spread and center when comparing groups, rather then comparing individual data points or ‘slices’ of the graphs.
- It is important to examine both the variability within a group (looking at how the data vary within one or more data sets) and the variability between groups (the variability of measures used to summarize and compare data sets), and distinguish between these two types of variability.

4. Recognizing variability in special types of distributions

- In a normal distribution, the mean and SD provide useful and specific information about variability. For example, if we know the SD in addition to the mean, we can determine the percentage of data within one, two, and three Standard Deviations of the mean. We can also use the mean and SD to estimate the high and low values of the distribution and the Range.
- There is variability in a bivariate data distribution, and we need to consider the variability of both variables as well as the variability for y values given individual values of x.
- The variability of a bivariate data set (covariation) may reveal a relationship between the variables and whether we might predict values of one variable (y) for values of the other (x).

5. Identifying patterns of variability in fitting models

- There is variability involved in fitting models and judging the fit of models (e.g., fitting the normal curve to a distribution of data, or fitting a straight line to a scatterplot of bivariate data).
- The variability of the deviations from the model (residuals) can tell us about the how well the model fits the data.
- Data may sometimes be reorganized and transformed to better reveal patterns or fit a model.

6. Using variability to predict random samples or outcomes

- Samples vary in some predictable ways, based on sample size and the population from which they are drawn and how they are drawn. If we have random samples the variability can be more readily explained and described.
- Larger samples have more variability than smaller samples, when randomly drawn from the same population. However, sample statistics from the larger samples vary less than statistics from smaller samples.
- There is variability in outcomes of chance events. We can predict and describe the variability for random variables.
- In some situations we can link the variability in samples to variability in outcomes, making predictions or statistical inferences.

7. Considering variability as part of statistical thinking

- In statistical investigations, we always need to begin with examining and discussing the variability of data.
- Data production is designed with variation in mind. Aware of sources of uncontrolled variation, we avoid self-selected samples, insist on comparison in experimental studies, and introduce planned variation into data production by use of randomization (Moore, 1990).
- In statistical analysis we try to explain variation by seeking the systematic effects behind the random variability of individuals and measurements (Moore, 1990).
- The ideas listed above are all part of statistical thinking, and come into play when exploring data and solving statistical problems (Wild & Pfannkuch, 1999).

While the ideas described above may seem like a list for a curriculum or syllabus of introductory statistics courses, we point out that in each area, the idea of variability is to be highlighted, discussed and emphasized. This list of increasingly sophisticated ideas offers: (1) the ways in which this body of knowledge can be structured so that it can be most readily grasped by the learner, (2) an effective sequences in which to present material related to variability, (3) a plan for revisiting variability as students progress through the statistics curriculum, and (4) a scaffold on which to build new levels of deep understanding of variability. Supported by theories of constructivist learning (e.g., Bruner, 1973; Cowan 1995), our belief is that progress in students' construction of meanings is not linear but rather complex and is better captured by the image of spiral progression. Therefore, ideas related to variability must be constantly revisited along the statistics curriculum from different points of view, context and levels of abstraction, to create a complex web of interconnections among them. An important part of building this complex knowledge is assessing where students are in developing their deep understanding and reasoning about variability, which is discussed in the next section.

3. ASSESSING STUDENTS' REASONING ABOUT VARIABILITY

Most traditional assessments regarding knowledge of variability focus on knowledge of definitions, calculations, and simple interpretations of standard measures of dispersion (e.g., computing and interpreting the Range, SD, and IQR). On the other hand, alternative assessments focus on intuitions, conceptual understanding, and reasoning about the connections between variability and other concepts such as center and shape. Most innovative assessment tasks in this area have been developed by researchers, as illustrated in the papers in these special issues. However, even such studies lack detailed discussions of how to assess students' knowledge and reasoning about variability. The ways students' knowledge has been revealed or evaluated in the two collections of studies on reasoning about variability is primarily through extended data investigation activities or tasks, and through open-ended survey questions. While these are found beneficial for research purposes, they are not as useful as assessment tasks to provide practical evaluation feedback to teachers and students in real classroom situations, in which time and teacher resources are limited.

It is essential to coordinate the desired learning outcomes in either research or classroom work with ways to assess these aspects of conceptual understanding (Wiggins & McTighe, 1999). Acknowledging that traditional methods of assessing understanding of variability have focused mostly on definitions, calculations, and simple interpretations of measures of spread, we reflected on the tasks used in the research articles reviewed and on their implications, as well as examined prior literature (e.g., Ben-Zvi & Garfield, 2004). This provided us with a basis for suggesting a broader array of assessment items to be used for evaluating students' reasoning about the complex concept of variability. In designing assessments that are aligned to instructional goals, we see a need to develop assessment items that involve a real or realistic context in order to motivate students and encourage them to use their informal knowledge of variability. Below we provide suggestions for developing assessment items, using real or realistic data, to evaluate student understanding of the many aspects of understanding variability. This list is organized in terms of the seven areas listed earlier in which students are expected to develop some knowledge or intuitions related to variability.

1. Assessment - Developing intuitive ideas of variability

- Items that provide descriptions of variables or raw data sets (e.g., the ages of children in a grade school, or the height of these children) and asking students to describe variability or shape of distribution.
- Items that ask students to make predictions about data sets that are not provided (e.g., if the students in this class were given a very easy test, what would you predict for the expected graph and expected variability of the test scores?).
- Given a context, students are asked to think of ways to decrease the variability of a variable (e.g., measurements of one students' jump).
- Items that ask students to compare two or more graphs and reason about which one would have larger or smaller measures of variability (e.g., Range or Standard Deviation).

2. Assessment - Describing and representing variability

- Items that provide a graph and summary measures, and ask students to interpret it and write a description of the variability for each variable.
- Items that ask students to choose appropriate measures of variability for particular distributions (e.g., IQR for skewed distribution) and select measure of center that are appropriate (e.g., median with IQR, mean with SD).
- Items that provide a data set with an outlier that ask students to analyze the effect of different measures of spread if the outlier is removed. Or, given a data set without an outlier, asking students what effect adding an outlier will have on measures of variability.
- Items that ask students to draw graphs of distributions for data sets with given center and spread.

3. Assessment - Using variability to make comparisons

- Items that present two or more graphs and ask students to make a comparison either to see if an intervention has made a difference or to see if intact groups differ. For example, asking students to compare two graphs to determine which one of two medicines is more effective in treating a disease, or whether there is a difference in length of first names for boys and girls in a class.
- Items that ask students which graph shows less (or more) variability, where they have to coordinate shape, center, and different measures of spread.

4. Assessment - Recognizing variability in special types of distributions

- Items that provide the mean and standard deviation for a data set that has a normal distribution and students are asked to use these to draw graphs showing the spread of the data.

- Items that provide a scatterplot for a specific bivariate data set and students have to consider if values are outliers for either the x or y variables or for both.
- Items that provide graphs of bivariate data sets where students are asked to determine if the variability in one variable (y) can be explained by the variability in the other variable (x).

5. Assessment - Identifying patterns of variability in fitting models

- Items that ask students to determine if a set of data appear normal, or if a bivariate plot suggests a linear relationship, based on scatter from a fitted line.

6. Assessment - Using variability to predict random samples or outcomes

- Items that provide students choices of sample statistics (e.g., proportions) from a specified population (e.g., colored candies) for a given sample size and ask which sequence of statistics is most plausible.
- Items that ask students to predict one or more samples of data from a given population.
- Items that ask students which outcome is most likely as a result of a random experiment when all outcomes are equally likely (e.g., different sequences of colors of candies)
- Items that ask students to make conjectures about a sample statistic given the variability of possible sample means.

7. Assessment - Considering variability as part of statistical thinking

- Items that give students a problem to investigate along with a data set, that requires them to graph, describe, and explain the variability in solving the problem.
- Items that allow students to carry out the steps of a statistical investigation, revealing if and how the students consider the variability of the data.

We suggest that teachers and researchers develop items along the lines suggested in the above list, embedding the tasks in contexts that provide real or realistic situations to engage students in reasoning about variability and to reveal their cognitive understanding of this concept. Such tasks will enhance students' ability to negotiate the complex and multifaceted meanings of variability suggested above. The results of these assessments should provide detailed pictures of what students understand or partially understand and how they reason about the many aspects of variability as they learn statistics and carry out exploratory data investigations. In the summary comments below we focus mainly on instructional implications of our deliberations.

4. IMPLICATIONS

We have enjoyed reading and reflecting on the rich collection of papers in the two special issues. As guest editors of these special issues we chose to extend the discussion to how we might build on the implications of the research papers in terms of teaching, curriculum, and assessing reasoning about variability. We hope that teachers of statistics at all levels will build on the results of these studies, finding and using good and rich data sets and contexts that motivate students to describe interpret, and reason about variability.

We proposed in this paper a model of statistical ideas that are part of a deep and multifaceted understanding and reasoning about variability. This epistemological model suggests a sequence of seven increasingly sophisticated areas of variability knowledge. This model can serve to guide and organize results from prior and future research. It can help educators and curriculum developers to plan and design instruction and assessment that aim at achieving deep understanding of variability, in the spirit of Perkins' (1993) "teaching and learning for understanding." We believe it is important to spend ample time on each area, with many examples in rich contexts. Variability should be centrally emphasized throughout the statistics curriculum, from the earliest grades (in formal and informal activities and discussions) through high school and the introductory college course. The areas of the model should be revisited often, making explicit the interconnections among them as well as the connections with other statistical "big ideas" (Garfield & Ben-Zvi, 2004).

Challenging students to reason about variability will require data sets in which variability is revealed in interesting ways, and lead students to make and test conjectures. We encourage teachers to conduct discussions on sources of variability and what leads to the variability in a data set rather than merely collecting data to analyze in graphs and summary measures, or presenting data to be summarized in this way. Activities and discussions that reveal the different aspects and uses of variability can begin in the early grade levels and should be repeatedly included as students progress in their study of statistics. In this way, intuitive, informal notions of variability can gradually be used to construct more formal and complex understanding of variability.

As pointed out in the discussion by Pfannkuch (2005, this issue) and by others (e.g., Garfield & Burrill, 1997), technological tools should play an important role both in developing students' reasoning about variability as well as in helping to make their reasoning visible to teachers and researchers. We recommend the careful use of appropriate technological tools to facilitate conceptual development and reasoning, and to assist in assessing and evaluating students' learning (e.g., Ben-Zvi, 2000). It seems clear from the studies in these special issues that reflection is an important component of these activities as well, and that students should be encouraged to communicate their learning and understanding in increasingly appropriate statistical language. We hope that assessments will be created and shared with teachers and researchers that carefully evaluate students' knowledge and reasoning about variability so that we may measure students' achievement of this important learning outcome. Finally, we emphasize the critical need for future research focused on the development of a deep understanding of variability and how such an understanding may be advanced over time and through the careful integration of curriculum, technology and assessment.

REFERENCES

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64-83. [Online: www.stat.auckland.ac.nz/serj]

Ben-Zvi, D. (2000). Towards understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2*(1&2), 127-155.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42-63. [Online: www.stat.auckland.ac.nz/serj]

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

Bruner, J. (1973). *Going beyond the information given*. New York: Norton.

Cowan, D. A. (1995). Rhythms of learning: Patterns that bridge individuals and organizations. *Journal of Management Inquiry, 4*(3), 222-246.

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

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

Garfield, J., & Burrill, G. (Eds.) (1997). *Research on the role of technology in teaching and learning statistics*. Voorburg, The Netherlands: International Statistical Institute.

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal, 3*(2), 7-16. [Online: www.stat.auckland.ac.nz/serj]

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/serj]

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

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, DC: National Academy Press.

Moore, D. (1992). Teaching statistics as a respectable subject. In F. Gordon & S. Gordon (Eds.), *Statistics for the twenty-first century*. MAA Notes, no. 26 (pp. 14-25). Washington, DC: Mathematical Association of America.

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

Perkins, D. (1993). An apple for education: Teaching and learning for understanding. *American Educator*, *17*(3), 8, 28-35.

Pfannkuch, M. (1997). Statistical thinking: One statistician's perspective. In F. Biddulph & K. Carr (Eds.), *People in mathematics education* (Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia) (pp. 406-413). Rotorua, New Zealand: MERGA.

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

Pfannkuch, M. (2005). Thinking tools and variation. *Statistics Education Research Journal, 4*(1), 83-91. [Online: www.stat.auckland.ac.nz/serj]

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

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

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

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 20-26.

Wiggins, G., & McTighe, J. (1999). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review, 67*(3), 223-265.

JOAN GARFIELD
University of Minnesota
315 Burton Hall
178 Pillsbury Drive, S.E.
Minneapolis, MN 55455, USA

PAST IASE CONFERENCES

1. IASE SATELLITE CONFERENCE - STATISTICS EDUCATION AND THE COMMUNICATION OF STATISTICS
Sydney, Australia, April 4-5, 2005

The conference was organized by the IASE and the Victorian Branch of the Statistical Society of Australia and jointly chaired by Brian Phillips, bphillips@swin.edu.au and Kay Lipson, klipson@swin.edu.au. Below is the listing of sessions and presented papers. Papers (in PDF format) are available at www.stat.auckland.ac.nz/~iase/publications.php?show=14.

Session I - Keynote Speaker

- To tell the truth: What we know about lie detection
  Stephen Fienberg

Session II - Writing Reports of Statistical Studies

- The implications introducing report writing into an introductory statistics subject
  Kay Lipson, Sue Kokonis
- Helping students find their statistical voices
  Helen MacGillivray
- There's more to statistics than computation - Teaching students how to communicate statistical results
  Roxy Peck
- An approach to report writing in statistics courses
  Glenda Francis
- Teaching students to write about statistics
  Mike Forster, David P. Smith, Christopher J. Wild
- Enabling students to communicate statistical findings
  Tania Prvan, Judith Ascione

Session III - Role of Graphics in Communication of Statistics

- Interval estimates for statistical communication: Problems and possible solutions
  Geoff Cumming, Fiona Fidler
- Using Amos Graphics to enhance the understanding and communication of multiple regression
  Everarda Cunningham, Wei C. Wang
- Grapharti
  Hilary Green
- Visualising data with dynamic graphics in Excel
  Harold Henderson
- Enhancing effective communication of statistical analysis to non-statistical audiences
  Peter Martin
- From data to graphs to words, but where are the models?
  K. Larry Weldon

Statistics Education Research Journal, 4(1), 100-110, http://www.stat.auckland.ac.nz/serj

Session IV - Outreach to Public and Schools

- Statistics and the media
  Wayne Smith
- The role of national statistics institutions in the use and understanding of official statistics in the compulsory education sector
  Gareth McGuinness, Lesliey Hooper
- Data analysis or how high school students "read" statistics
  Sylvain Vermette, Linda Gattuso, Marc Bourdeau
- Communicating student performance data to school teachers
  Philip Holmes-Smith
- Pilot study into the use and usefulness of instant messaging within an educational context
  Rachel Cunliffe
- Distance learning: New frontiers for solving old problems
  Gianfranco Galmacci, Anna Maria Milito
- Helping students to communicate statistics better
  Neville Davies, Doreen Connor

Session V - Understanding the Language of Statistics

- Servicing students communicating ideas about statistics
  Peter Petocz, Anna Reid
- How important are communication skills for "good" statistics students? - An international perspective
  Sue Gordon
- University-level data analysis courses with the emphasis on understanding and communication of statistics - A ten years action research project
  Katrin Niglas, Kairi Osula
- Statistical prevarication: Telling half truths using statistics
  Milo Schield

2. THE 2005 SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE, ISI-55
Sydney, Australia, April 5-12, 2005

1.1 IASE ACTIVITIES AT THE ISI-55

IASE organized a wide and varied list of topics at the 55th Session of the International Statistical Institute for Invited Paper Meetings, both alone and in conjunction with other ISI sections and Committes. The Chair of IASE Programme was Chris Wild, c.wild@auckland.ac.nz. Below is the listing of the Invited Paper Meetings, organizers and presented papers. Papers (in PDF format) are available at www.stat.auckland.ac.nz/~iase/publications.php?show=13

IPM 45 *Reasoning about Variation*. Christine Reading, creading@une.edu.au

- From acknowledging to modelling: Tertiary students' consideration of variation
  Jackie Reid, Chris Reading
- Some aspects of reasoning about variability
  Bernard Harris
- Statistical thinking from a practitioner's perspective
  Aloke Phatak, Geoff Robinson

IPM 46 *The use of Simulation in Statistics Education*. Andrej Blejec, andrej.blejec@nib.si

- Modern introductory statistics using simulation and data analysis
  Larry Weldon
- Using Excel to generate empirical sampling distributions
  Rodney Carr, Scott Salzman
- Statistical simulations in the web
  Juha Puranen

IPM 47 *Teaching Statistics Online*. Larry Weldon, weldon@sfu.ca

- Learning statistics teaching in higher education using online and distance methods
  Neville Davies, Vic Barnett
- Preparing secondary teachers to teach statistics: A distance education model
  Roxy Peck, Robert Gould
- E-learning for statistics education at Korea National Open University
  Tae Rim Lee

IPM 48 *Statistics for Life: What are the Statistical Ideas or Skills that Matter most and why?* Chris Wild, c.wild@stat.auckland.ac.nz

Panel participants:
Nick Fisher, Denise Lievesley, Milo Schield, Stephen Stigler, Niels Keiding

IPM 49 *Research in Statistical Education*. Kay Lipson, klipson@swin.edu.au / Maria Ottaviani, Mariagabriella.ottaviani@uniroma1.it

- An assessment of computer-based learning methodology in teaching in an introductory statistics hybrid course
  Paul Fields, Patti Collins
- Potential uses of longitudinal analyses to investigate statistics education outcomes
  Sharleen Forbes, Teimuraz Beridze
- Teaching confidence intervals: Problems and potential solutions
  Geoff Cumming, Fiona Fidler
- Student opinions and expectations vs. reality of grading: Use of cluster profiling in statistics education
  Mojca Bavdaz-Kveder, Irena Ograjensek

IPM 50 *Quality Assurance in Statistics Education*. Matthew Regan, m.regan@auckland.ac.nz

- Quality assurance in statistics education: From departmental self-evaluation to accreditation
  Abbas Bazargan
- The role of statistical education in developing graduate qualities
  Brenton Dansie
- Criteria, standards and assessment in statistical education
  Helen MacGillivray

IPM 51 *Promotion of Statistical Literacy among Students*. Pilar Guzman, pilar.guzman@uam.es

- The role of official statistics agencies in the promotion of statistical literacy among students
  Frederick W. H. Ho
- Co-operation with educational institutions: A strategic challenge for statistical agencies
  Reija Helenius
- Policies and tools to make OECD statistics more visible and accessible
  Enrico Giovannini, Russell Penlington, Lars Thygese

IPM 52 *Using History of Statistics to Enhance the Teaching of Statistics*. (IASE & Christiaan Huygens Com. on the History of Statistics) Carol J. Blumberg, cblumberg@winona.edu

- Probability and statistics ideas in the classroom - Lessons from history
  David Bellhouse
- Taking the fear out of data analysis: Case for history lessons in statistics courses
  Irena Ograjensek
- Teaching probability via its history: Reflections on a case study
  David Vere-Jones

IPM 63 *Educating the Media on how best to Report Statistics*. Jacob Ryten, rytenjacob@msn.com

- Introduction to media and official statistics
  Jacob Ryten
- Panel presentations and floor discussion
  Peter Harper, Ross Gittens, Tim Colebatch, Frederick W. H. Ho, Beat Hulliger

IPM 81 *Ethical Standards in Statistics Education*. (IASE & ISI Committee on Professional Ethics) Mary A. Gray, mgray@american.edu

- Making a difference, not faking a difference - Learning And Using What's Good And Fair In Biostatistics
  David Goddard
- The client-consultant relationship in medical research: The role of a professional statistician in the research team
  Nora Donaldson, Mary Gray
- Official statistics and statistical ethics: Selected issues
  William Seltzer

IPM 83 *Challenges in the Teaching of Survey Sampling.* (IASS & IASE) Wilton de Oliveira Bussab, bussab@fgvsp.br

- Balancing statistical theory, sampling concepts, and practicality in the teaching of survey sampling
  Colm O'Muircheartaigh
- Teaching environment for survey sampling based on a textbook and its web extension
  Risto Lehtonen
- Teaching sampling in a government statistical agency: The Canadian experience
  Jack Gambino, Hew Gough

OTHER PAST CONFERENCES

1. ASIAN TECHNOLOGY CONFERENCE IN MATHEMATICS 2004
Singapore, December 13-17, 2004

The conference program (with links to abstracts) of the Asian Technology Conference in Mathematics on the theme Technology in Mathematics: Engaging Learners, Empowering Teachers and Enabling Research is available at

www.atcminc.com/mConferences/ATCM04/TentativeConferenceProgram.html

Conference web page: www.atcminc.com/mConferences/ATCM04/

2. CONGRESS OF THE EUROPEAN SOCIETY FOR RESEARCH IN MATHEMATICS EDUCATION, CERME-4
Sant Feliu de Guíxols, Spain, February 17-21, 2005

The Fourth Congress of the European Society for Research in Mathematics Education (CERME) was held in Sant Feliu de Guíxols, Spain, from 17 to 21 February, 2005. Presented papers (in PDF format) are available at http://cerme4.crm.es/5.htm.

Of special interest are papers presented in Group 5: Stochastic Thinking (subjects include epistemological and educational issues, pupils cognitive processes and difficulties, and curriculum issues), organized by Dave Pratt (e-mail: dave.pratt@warwick.ac.uk), Rolf Biehler, Maria Gabriella Ottaviani, and Maria Meletiou. The papers presented in Group 5 can be found at cerme4.crm.es/Papers%20definitius/5/wg5listofpapers.htm

FORTHCOMING IASE CONFERENCES

1. SRTL-4 THE FOURTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING AND LITERACY

Auckland, New Zealand, July 2-7, 2005

The Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy, is to be hosted by the Department of Statistics, The University of Auckland, New Zealand, July 2–7, 2005. This gathering offers an opportunity for a small, interdisciplinary group of researchers from around the world to meet for a few days to share their work, discuss important issues, and initiate collaborative projects. Having emerged from the three previous forums, the topic and focus of SRTL-4 will be Reasoning about Distribution. The Forum is co-chaired by Dani Ben-Zvi (University of Haifa, Israel) and Joan Garfield (University of Minnesota, USA), co-organized by Maxine Pfannkuch and Chris Wild (The University of Auckland, New Zealand), and planned by a prestigious international advisory committee.

Based on the SRTL tradition, we plan to keep the number of participants small to facilitate a working research forum. There are three possible roles for participants in this Forum. The first role is to present current research on reasoning about distribution, the second is to discuss and react to research presentations, while the third is to be a small group moderator, which is ideal for doctoral students who are not yet ready to present research but want to participate. Participants will be strongly encouraged to use videotape and written transcripts of students in classroom and interview settings to provide illustrations of what the researchers are learning about how students reason about distribution. As with the previous SRTL Research Forums, we encourage the participation of young promising scholars. One outcome of the Forum will be a publication summarizing the work presented, discussions conducted, and issues emerging from this gathering.

The SRTL-4 Research Forum organizers invite anyone interested in participating in this forum to contact them as soon as possible. More information from Maxine Pfannkuch, m.pfannkuch@auckland.ac.nz. Web site: www.stat.auckland.ac.nz/srtl4/

2. 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.

The major aim of ICOTS-7 is to provide the opportunity for people from around the world who are involved in statistics education to exchange ideas and experiences, to discuss the latest developments in teaching statistics and to expand their network of statistical educators. The conference theme emphasises the idea of *cooperation,* which is natural and beneficial for those involved in the different aspects of statistics education at all levels.

2.1. CALL FOR PAPERS

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

Voluntary refereeing procedures will be implemented for ICOTS7. Details of how to prepare manuscripts, the refereeing process and final submission arrangements will be announced later.

Invited Papers

Invited Paper Sessions are organized within 9 different Conference Topics 1 to 9. The list of Topic and Sessions themes, with email contact for Session Organisers is available at the ICOTS-7 web site at www.maths.otago.ac.nz/icots7, under "Scientific Programme".

Contributed Papers

Contributed paper sessions will be arranged in a variety of areas. Those interested in submitting a contributed paper should contact either Joachim Engel (Engel_Joachim@ph-ludwigsburg.de) or Alan MacLean (alan.mclean@buseco.monash.edu.au) before September 1, 2005.

Posters

Those interested in submitting a poster should contact Celi Lopes (celilopes@uol.com.br) before February, 1, 2006.

Special Interest Group Meetings

These are meetings of Special Interest Groups of people who are interested in exchanging and discussing experiences and/ or projects concerning a well-defined theme of common interest. Proposals to hold a SIG Meeting specifically oriented to reinforce Latin American statistics education cooperation in a particular theme are especially welcome. In this case the organisers may decide to hold the meeting in Portuguese and Spanish language. Individuals or groups may submit proposals to establish a Special Interest Group to Carmen Batanero at (batanero@ugr.es).

2.2. TOPICS AND TOPIC CONVENORS

Topic 1. *Working Cooperatively in Statistics Education*. Lisbeth Cordani, lisbeth@maua.br and Mike Shaughnessy, mike@mth.pdx.edu

Topic 2. *Statistics Education at the School Level*. Dani Ben-Zvi, benzvi@univ.haifa.ac.il and Lionel Pereira, lpereira@nie.edu.sg

Topic 3. *Statistics Education at the Post Secondary Level*. Martha Aliaga, martha@amstat.org and Elisabeth Svensson, elisabeth.svensson@esi.oru.se

Topic 4. *Statistics Education/Training and the Workplace*. Pedro Silva, pedrosilva@ibge.gov.br and Pilar Martín, pilar.guzman@uam.es

Topic 5. *Statistics Education and the Wider Society*. Brian Phillips, BPhillips@groupwise.swin.edu.au and Phillips Boland, Philip.J.Boland@ucd.ie

Topic 6. *Research in Statistics Education*. Chris Reading, creading@metz.une.edu.au and Maxine Pfannkuch, pfannkuc@scitec.auckland.ac.nz

Topic 7. *Technology in Statistics Education.* Andrej Blejec, andrej.blejec@nib.si and Cliff Konold, konold@srri.umass.edu

Topic 8. *Other Determinants and Developments in Statistics Education*. Theodore Chadjipadelis, chadji@polsci.auth.gr and Beverley Carlson, bcarlson@eclac.cl

Topic 9. *An International Perspective on Statistics Education*. Delia North, delian@icon.co.za and Ana Silvia Haedo, haedo@qb.fcen.uba.ar

Topic 10. *Contributed Papers*. Joachim Engel, Engel_Joachim@ph-ludwigsburg.de and Alan McLean, alan.mclean@buseco.monash.edu.au

Topic 11. *Posters*. Celi Espasandín López, celilopes@directnet.com.br

2.3. ORGANISERS

Local Organisers

Pedro Alberto Morettin, (Chair; pam@ime.usp.br), Lisbeth K. Cordani (lisbeth@maua.br), Clélia Maria C. Toloi (clelia@ime.usp.br), Wilton de Oliveira Bussab (bussab@fgvsp.br), Pedro Silva (pedrosilva@ibge.gov.br).

IPC Executive

Carmen Batanero (Chair, batanero@ugr.es), Susan Starkings (Programme Chair, starkisa@lsbu.ac.uk), Allan Rossman and Beth Chance (Editors of Proceedings; arossman@calpoly.edu; bchance@calpoly.edu), John Harraway (Scientific Secretary: jharraway@maths.otago.ac.nz), Lisbeth Cordani (Local organisers representative; lisbeth@maua.br).

More information is available from the ICOTS-7 web site at www.maths.otago.ac.nz/icots7 or from the ICOTS IPC Chair Carmen Batanero (batanero@ugr.es), the Programme Chair Susan Starkings (starkisa@lsbu.ac.uk) and the Scientific Secretary John Harraway (jharraway@maths.otago.ac.nz).

3. 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. Preliminary information can be found at www.tziranda.com/isi2007/

OTHER FORTHCOMING CONFERENCES

1. UNITED STATES CONFERENCE ON TEACHING STATISTICS, USCOTS
Columbus, OH, USA, May 19-21, 2005

The first United States Conference on Teaching Statistics (USCOTS) will be held on May 19-21, 2005 at the Ohio State University in Columbus, Ohio, hosted by CAUSE, the Consortium for the Advancement of Undergraduate Statistics Education. USCOTS is an active, hands-on working conference for teachers of Statistics at the undergraduate level, in any discipline or type of institution, including high school teachers of AP Statistics. USCOTS will feature spotlight sessions, plenary talks, and working breakout sessions in three major areas: curriculum, pedagogy, and research. Lots of good resources for each of these areas will be provided in a fun and active atmosphere, where everyone will be invited to be involved. The theme of the 2005 USCOTS is "Building Connections for Undergraduate Statistics Teaching" and will focus on ways that we can share teaching ideas, develop working relationships, and identify areas for future collaborations and projects at our own institutions. USCOTS is partially funded by The Ohio State University Department of Statistics and its College of Mathematical and Physical Sciences. For more information about USCOTS, please contact Deborah Rumsey, USCOTS program chair at rumsey@stat.ohio-state.edu. Web site: www.causeweb.org/uscots/

2. INTERNATIONAL CONFERENCE ON MATHEMATICAL MODELLING AND APPLICATIONS, ICTMA 12
London, United Kingdom, July 10-14, 2005

Mathematical modelling and applications, the transition, freely between real world problems and mathematical representations of such problems, is an enduring and important feature of industry, business and commerce. Teaching mathematical modelling, through tasks, projects investigations and applications embedded in courses and of mathematics itself through applications helps learners to understand relationships between real world problems and mathematical models. The 12th International Conference on Mathematical Modelling and Applications (ICTMA12) will be hosted by the School of Engineering and Mathematical Sciences City of London, Sir John Cass Business School, London, UK. This conference brings together international experts in a variety of fields and from different sectors to consider: modelling in business and industry, evaluating effectiveness, pedagogic issues for learning, applicability at different levels, research: education and practice, innovative practices and transitions to expert practice. More information from Chris Haines, ictma12@city.ac.uk. Web site: www.city.ac.uk/conted/reseach/ictma12/index.htm

3. PSYCHOLOGY OF MATHEMATICS EDUCATION, PME-29
Melbourne, Australia, July 10-15, 2005

The PME-29 conference will be held on July 10-15, 2005 in Melbourne, Australia. More information from Helen Chick, h.chick@unimelb.edu.au.

Web site: staff.edfac.unimelb.edu.au/~chick/PME29/

4. THE 25TH EUROPEAN MEETING OF STATISTICIANS, EMS 2005
Oslo, Norway, July 24-28, 2005

The meeting will cover all areas of methodological, applied and computational statistics, probability theory and applied probability. There will be 8 special lecturers, 23 ordinary invited sessions and one invited discussion session. The scientific programme is broad, with ample space for applications; invited speakers and sessions have been chosen with the specific aim to appeal to a wide audience and will bridge between theory and practice, inference and stochastic models. The meeting is organised jointly by the University of Oslo and the Norwegian Computing Center. For further information contact email to ems2005@nr.no or visit the web site: www.ems2005.no

5. BEYOND THE FORMULA IX
Rochester, NY, USA, August 4-5, 2005

Beyond The Formula is an annual two-day, summer conference designed to promote excellence in teaching introductory statistics. Whether participants come from the high school, two-year or four-year college setting, they can expect to hear speakers who will provide them with new ideas and techniques that will make their classrooms more effective statistics learning centers. In organizing BTF conferences four major areas of teaching concern have been identified: the curriculum, the techniques and methodologies, the ever-changing technology, and real world applications. Each year one of these topic areas is chosen to serve as a common thread to draw together all the presentations - with 20 to 25 large-group, small-group and workshop sessions, all of the four topic areas find their way into each conference. Come with us each summer on a trip Beyond The Formula to see how exciting, thought provoking and inspiring the teaching of statistics can really be. For more information contact Robert Johnson, rr.bs.johnson@juno.com. Web site: www.monroecc.edu/go/beyondtheformula/

6. JOINT STATISTICAL MEETINGS, JSM 2005
Minneapolis, MN, USA, August 7-11, 2005

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 4000 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), placement service, society and section business meetings, committee meetings, social activities, and networking opportunities. Minneapolis is the host city for JSM 2005 and offers a wide range of possibilities for sharing time with friends and colleagues. For information, contact Elaine Powell, jsm@amstat.org or phone toll-free (800) 308-8943. Web site: www.amstat.org/meetings/jsm/2005/

7. 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 conference will be organized in Johor Bharu, Malaysia, close to Singapore and in the heart of Tropical South East Asia. Mercifully this part of Malaysia was unaffected by the tragic earthquake and tidal wave at the end of 2004. The time and place were chosen to encourage teachers and mathematics educators from around the world to communicate with each other about the challenges of Reform, Revolution and Paradigm Shifts in Mathematics Education. The Malaysia Conference is organised by the Mathematics Education into the 21st Century Project - an international educational initiative whose coordinators are Dr. Alan Rogerson (Poland) and Professor Fayez Mina (Egypt). Since its inception in 1986, the Mathematics Education into the 21st Century Project has received support and funding from many educational bodies and institutions throughout the world. In 1992 UNESCO published our Project Handbook "Moving Into the 21st Century" as Volume 8 in the UNESCO series Studies In Mathematics Education.

The Mathematics Education into the 21st Century Project is dedicated to the improvement of mathematics education world-wide through the publication and dissemination of innovative ideas.

Many prominent mathematics educators have supported and contributed to the project, including the late Hans Freudental, Andrejs Dunkels and Hilary Shuard, as well as Bruce Meserve and Marilyn Suydam, Alan Osborne and Margaret Kasten, Mogens Niss, Tibor Nemetz, Ubi D'Ambrosio, Brian Wilson, Tatsuro Miwa, Henry Pollack, Werner Blum, Roberto Baldino, Waclaw Zawadowski, and many others throughout the world. For further conference details email to Alan Rogerson, arogerson@vsg.edu.au. Website: math.unipa.it/~grim/21_malasya_2005.doc

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

The 10^{th} Annual conference of Asian Technology Council in Mathematics (ATCM) on the theme Enriching Technology in Enhancing Mathematics for All is hosted by the Korea National University of Education in Cheong-Ju, South Korea. The aim of this conference is 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. English is the official language of the conference. The conference will cover a broad range of topics on the application and use of technology in Mathematics research and teaching. More information: Professor Hee-chan Lew, hclew@knue.ac.kr. Website: www.atcminc.com/mConferences/ATCM05/