**Paper Title** Productive Failure in Learning the Concept of Variance

**Author(s)** Manu Kapur, National Institute of Education, Singapore; June Lee, National Institute of Education, Singapore

**Session Title** Exploring Student Thinking, Reasoning, and Conceptual Development Across the Mathematics Curriculum

**Session Type** Paper

**Presentation Date** 5/3/2010

**Presentation Location** Denver, Colorado, USA

**Descriptors** Mathematics Education, Classroom Research, Instructional Interventions

**Methodology** Mixed Method

**Unit** Division C - Learning and Instruction

Productive Failure in Learning the Concept of Variance

ABSTRACT

We report on a quasi-experimental study comparing a "*productive failure (PF)*" design with a "*direct instruction (DI)*" design for a curricular unit on variance. $N = 140$, $9^{th}$-grade mathematics students from an all-boys secondary school in Singapore experienced either DI or a PF design, where they solved a complex problem in small groups without the provision of any support up until a teacher-led consolidation. PF students produced a diversity of progressively sophisticated problem representations and methods for solving the problem but were ultimately unsuccessful in developing the canonical solution. Despite seemingly failing in their problem-solving efforts, PF students performed on par with DI students on well-structured problems on variance, and significantly outperformed them on complex, data analysis problems.

We report on an on-going design research program on *productive failure* (Author, 2008) in mathematical problem solving. The focus is on designing conditions for learners to persist in the process of solving complex problems without instructional support structures initially. Even though such a process invariably leads to failure in the shorter term, we argue that the extent to which our design affords learners opportunities to persist in exploring and generating a variety of representations and methods for solving the problem, the process can be germane for learning.

Indeed, there is a growing body of research that suggests that conditions that maximize performance in the shorter term are not necessarily the ones that maximize learning in the longer term (Clifford, 1984; Schmidt & Bjork, 1992). Examples of such research programs include VanLehn's (2003) work on *impasse-driven learning*, Schwartz and Bransford's (1998) work on *preparation for future learning*, Schwartz and Martin's (2004) work on *inventing to prepare for learning*, diSessa's (1991) work on *meta-representational competence*, Lesh and Doerr's (2003) work on *model-eliciting activities*, Koedinger and Aleven's (2007) work on the *assistance dilemma*, among others (e.g. Even, 1998; Dixon & Bangert, 2004; Slamecka & Graf, 1978).

Word limit constraints do not permit a fuller theoretical and empirical analysis of the abovementioned programs; we present such an analysis elsewhere (see Author, 2008, 2009; Author & Kinzer, 2009). For the present purposes, it suffices to reinterpret their central findings collectively as an argument to explore and design conditions for leveraging the hidden efficacy of learner-generated conceptions, representations, and understandings, even though such conceptions, representations, and understandings may not be correct initially and the process of arriving at them not as efficient.

## DESIGNING FOR PRODUCTIVE FAILURE

The purpose of this paper is to report findings from a classroom-based research study on productive failure in mathematical problem solving at a public school in Singapore. The targeted curricular unit was on statistical data analysis focusing specifically on the concept of variance.

### *Participants*

Participants were 140, Secondary 3 ($9^{th}$-grade) students from an all-boys secondary school in Singapore. Students were from four intact math classes; three classes taught by one teacher (teacher A), and the fourth class by another teacher (teacher B). Students had no instructional experience with the targeted concept—variance—prior to the study.

### *Research Design*

A pre-post, quasi-experimental design was used with two classes ($n$ = 31 and 35) taught by teacher A assigned to the 'Direct Instruction' (DI) condition, and the other two classes ($n$ = 35 and 39), under teachers A and B, assigned to the 'Productive Failure' (PF) condition. All classes participated in five lessons totaling 5 hours of instructional time over two weeks. There was no significant difference between the classes on the 30-minute, 5-item pre-test ($\alpha = .75$) pre-test, $F(3,136) = 1.665$, $p = .177$.

*PF Condition*. Students worked face-to-face in triads for two periods on a complex problem (see Appendix A) that afforded opportunities for generating multiple representations and solution methods. No extra support or scaffolds were provided during the group or individual problem-solving nor was any homework assigned at any stage. Following this, three periods were spent on consolidation where the teacher led a discussion of the targeted concept. Validation of the problem scenario was carried out through multiple iterations of design and pilot-testing.

*DI Condition*. Students were involved in teacher-led lectures, worked examples, and practice. The teacher introduced the concept of variance to the class, worked out examples, encouraged students to ask questions, following which students solved problems for practice. The teacher then discussed the solutions with the class. For homework, students were asked to continue with the workbook problems. This cycle of lecture, practice/homework, and feedback then repeated itself over the course of five periods. Note that the worked-out examples and practice problems were typically well-structured problems with fully-specified parameters, prescriptive representations, predictive sets of solution strategies and solution paths, often leading to a single correct answer.

In short, the DI condition represented a design that was highly structured from the very beginning with the teacher leading the students through a set of well-structured problems with proximal feedback and regular practice. The PF condition represented a design that delayed structure up until students had completed solving the complex problem without any instructional facilitation, support structures, or scaffolds. It is important to note that the research design allows for a comparison between instructional designs as wholes, not their constituent elements. Unlike laboratory experiments, the reality of classroom-based research is that one is rarely able to isolate individual elements of an instructional design in a single study because it is the complexity of how the individual elements combine that gives rise to the efficacy of a particular design (Brown, 1992).

*Hypothesis*. Based on past research on productive failure in Singapore schools (Author, 2009), we hypothesized that compared to the DI condition, the PF condition may result in students exploring various representations and methods for solving the complex problem (diSessa et al., 1991; Spiro et al., 1992). We did not expect students who were novices to the targeted concept of average speed to use the most effective representations and domain-specific methods for solving the problem, nor did we expect them to be successful in their problem-solving efforts (Chi et al., 1988; Kirschner et al., 2006). However, such a process may be integral to engendering the necessary knowledge differentiation which may help students better discern and understand those very concepts, representations, and methods when presented in a well-assembled, structured form during the consolidation lecture (Marton, 2007; Schwartz & Bransford, 1998).

### *Data Sources*

Process data included group work artifacts produced on A4 sheets of paper, audio-taped and transcribed group discussions as well as classroom videos (not reported here). Outcome data included performance on a posttest comprising six open-response items: two well-structured items and four complex, data analysis items ($\alpha$ = .74; see Appendix B for examples). Aggregate score on the two well-structured items and four complex, data analysis items formed the two dependent variables in our analysis. Pretest score was used as a covariate.

## RESULTS

### *Process Results*

Group-work artifacts provided a rich source of data about the nature of problem representations and methods generated by the groups in the process of solving the problem. Qualitative content analysis revealed that groups produced four major and *progressively sophisticated* categories of methods and representations for solving the problem. The four categories were: a) central tendencies, b) qualitative methods, c) frequency methods, and d) deviation methods.

*Category 1: Central Tendencies*. Groups started by using mean, media, and in some cases, mode for data analysis. The problem scenario was designed in a way that the mean and median were the same for all the three data sets. One data set was bi-modal and the other two had the same mode. Therefore, by relying on central tendencies alone, it was not possible to solve the problem; the design of the problem "forced" groups to consider alternative methods for generating an index for consistency.

*Category 2: Qualitative methods*. Figure 1 presents examples of qualitative methods generated by the groups. Groups generated graphical and tabular representations that organized the data visually and were able to discern which striker was more consistent. The visual representations afforded a qualitative comparative analysis between the players. However, these representations did not provide a quantitative index for measuring consistency even though the ideas of spread and clustering are quite evidently important qualitative conceptual underpinnings for the concept of variance.

*Figure 1*

*Category 3: Frequency methods*. Groups built on the qualitative methods to develop frequency-based measures of consistency. For example in Figure 2, groups used the frequency of goals scored within certain intervals to argue that the player with the highest number of goals in the interval containing the mean was the most consistent. Other groups counted the frequency with which a player scored above, below, and at the mean. Frequency methods demonstrated that could quantify the clustering and bunching up trends that the qualitative representations revealed. However, frequency-based methods ignore the relative distance between data points, and as such are sub-optimal as measures of consistency.

*Figure 2*

*Category 4: Deviation methods*. Figure 3 presents examples of the deviation methods. The simplest deviation method was the range. Some groups calculated year-on-year deviation as a percentage to argue that the greater the sum of the percentages, the lower the consistency. In the same vein, other groups used the sum of year-on-year deviations. Among these, there were those who considered absolute deviations to avoid deviations of opposite signs cancelling each other—an important conceptual leap towards understanding variance. Finally, there were some groups who calculated deviations about the mean only to find that they sum to zero. Others realized that they had to consider the absolute value of the deviations about the mean.

*Figure 3*

*Summary*. In both the PF classes, all groups demonstrated representational competence at the Category 3 level or greater. Only 2 groups from PF-A and 1 group from PF-B did not reach Category 4. Therefore, it is reasonable to suggest that in spite of attempting various representations and methods for solving the problem, PF students were ultimately unsuccessful in generating the canonical index for consistency. These process findings serve as a manipulation check demonstrating that students in the PF condition experienced "failure" at least in the conventional sense. In contrast, students in the DI condition, by design, repeatedly experienced performance success in solving well-structured problems under the teacher's close monitoring, scaffolding, and feedback.

*Outcome Results*

Controlling for the effect of prior knowledge, a MANCOVA revealed a statistically significant multivariate effect of condition (PF vs. DI) on posttest scores, $F(2, 136) = 14.55$, $p < .001$, partial $\eta^2 = .18$. There was no significant difference between the classes *within* the PF or DI conditions.

*Figure 4*

Univariate analysis further suggested that (Figure 4):

i. On the two well-structured items (maximum score = 10), there was no significant difference between the PF condition, $M = 6.78$, $SD = 2.71$, and the DI condition, $M = 6.98$, $SD = 2.51$, $F(1, 137) = .712$, $p = .400$. It is important to note that PF students who were not given any homework with minimal emphasis on practice still managed to perform on par with DI students who substantially received such homework, practice and feedback on well-structured types of items.

ii. On the four complex data analysis items (maximum score = 45), students from the PF condition, $M = 29.64$, $SD = 9.38$, significantly outperformed those from the DI condition, $M = 22.59$, $SD = 10.18$, $F(1, 137) = 15.219$, $p < .001$, partial $\eta^2 = .10$.

## DISCUSSION

Consistent with our previous findings in Singapore schools (Author, 2009; Author & Bielaczyc, under review), findings suggested that PF students outperformed those from the DI condition on the complex, data analysis items on the posttest without compromising performance on the well-structured items. As hypothesized, our analysis revealed that PF students explored and generated a diversity of representations and methods for solving the complex problems. They were evidently not successful, but the process of exploring the problem and solution spaces may have engendered sufficient knowledge differentiation that prepared them to better discern and understand those very concepts, representation, and methods when presented in a well-assembled, structured form during the consolidation lecture (Marton, 2007; Schwartz & Bransford, 1998; Schwartz & Martin, 2004).

It is of course much too early to attempt any generalization of the claims from two studies; the scope of inference technically holds only under the conditions and settings of the study. Going forward, therefore, future research would do well to extend this study to larger samples across schools and subjects. At the same time, further analyses of group discussions should unpack learning mechanisms underpinning the productive failure effect. In particular we will be analyzing variation within the PF condition to examine variation between groups in terms of their collaborative dynamics as well as the fidelity of the teacher-led consolidation discussions.

REFERENCES


Anderson, J. R. (2000). *Cognitive psychology and its implications*. New York, NY: Worth.

Author (2009). Productive failure in mathematical problem solving. *Instructional Science*. doi: 10.1007/s11251-009-9093-x.

Author (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379-424.

Author, & Bielaczyc, K. (under review). Design experiments in productive failure. Manuscript submitted for publication.

Author, & Kinzer, C. (2009). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning* (*ij*CSCL), *4*(1), 21-46.

Brown, A. L. (1992). Design experiments. *Journal of the Learning Sciences, 2*(2), 141-178.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum

Clifford, M. M. (1984). Thoughts on a theory of constructive failure. *Educational Psychologist*, *19*(2), 108-120.

diSessa, A. A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: meta-representational expertise in children. *Journal of Mathematical Behavior*, *10*(2), 117–160.

Dixon, J. A., & Bangert, A. S. (2004). On the spontaneous discovery of a mathematical relation during problem solving. *Cognitive Science*, *28*, 433-449.

Even, R. (1998). Factors involved in linking representations of functions. *Journal of Mathematical Behavior, 17*(1), 105-121.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist*, *41*(2), 75-86.

Koedinger, K. R., & Aleven V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*(3), 239-264.

Lesh, R. R., & Doerr, H. M. (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, NJ: Erlbaum.

Marton, F. (2007). Sameness and difference in transfer. *The Journal of the Learning Sciences*, *15*(4), 499-535.

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist*, *40*(1), 1-12.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207-217.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475-522.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129-184.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592-604.

Spiro, R. J., Feltovich, R. P., Jacobson, M. J., & Coulson, R. L. (1992). Cognitive flexibility, constructivism, and hypertext. In T. M. Duffy, & D. H. Jonassen (Eds.), *Constructivism and the technology of instruction: A conversation*. NJ : Erlbaum.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209-249.

## APPENDIX A: The Complex Problem Scenario

Mr. Fergusson, Mr. Merino, and Mr. Eriksson are the mangers of the Supreme Football Club. They are on the lookout for a new striker, and after a long search, they short-listed three potential players: *Mike Arwen*, *Dave Backhand*, and *Ivan Right*.

All strikers had asked for the same salary, so the managers agreed that they should base their decisions on the players' performance in the Premier League for the last 20 years. Table 1 shows the number of goals that each striker had scored between 1988 and 2007.

Mr. Fergusson calculated the average number of goals scored by each player over the past 20 years. He argued, "I think we should hire Ivan Right. He had the most number of years where he scored *more* goals than his average. This means that he pushes himself to score beyond his potential, and we need players like that".

"I disagree!" argued Mr. Merino, "Ivan was also the one with the most number of years with goals *fewer* than his average. I prefer Dave Backhand, who has the fewest number of years where he scored below his average. He seems to be a steadier performer, and we should hire him instead."

"Hm, I disagree with the both of you", argued Mr. Eriksson, "Dave may be good, but I think Mike is better. His performance over the years looks more consistent. He is my choice".

Although the three managers could not decide on whom to hire, they agreed that the player they hire should be a consistent performer.

*Table 1: Number of Goals scored by three potential strikers during the Premier League between 1988 and 2007.*

| Year | Mike Arwen | Dave Backhand | Ivan Right |
|---|---|---|---|
| 1988 | 14 | 13 | 13 |
| 1989 | 9 | 9 | 18 |
| 1990 | 14 | 16 | 15 |
| 1991 | 10 | 14 | 10 |
| 1992 | 15 | 10 | 16 |
| 1993 | 11 | 11 | 10 |
| 1994 | 15 | 13 | 17 |
| 1995 | 11 | 14 | 10 |
| 1996 | 16 | 15 | 12 |
| 1997 | 12 | 19 | 14 |
| 1998 | 16 | 14 | 19 |
| 1999 | 12 | 12 | 14 |
| 2000 | 17 | 15 | 18 |
| 2001 | 13 | 14 | 9 |
| 2002 | 17 | 17 | 10 |
| 2003 | 13 | 13 | 18 |
| 2004 | 18 | 14 | 11 |
| 2005 | 14 | 18 | 10 |
| 2006 | 19 | 14 | 18 |
| 2007 | 14 | 15 | 18 |

They decided that they should approach this decision mathematically, and would want a formula for calculating the consistency of performance for each player. This index should apply to all players and help provide a fair comparison.

The managers decided to get your help. **Please come up with a formula for consistency and show which player is the most consistent striker. You should make use of all data points to come up with the formula. Show all working and calculations on the paper provided.**

## APPENDIX B: Examples of Posttest Items

*Example of a Well-structured Item*

Marks scored by 10 students on a test on statistics are shown below. As a measure of the variance, calculate the *standard deviation* of the test scores.

30, 50, 50, 55, 60, 60, 60, 70, 80, 90

*Example of a Complex, Data Analysis Item*

In preparing for the Youth Olympics in 2010, the Ministry of Community, Youth and Sports had to decide the month in which to hold the games. They narrowed their options to July and August, and decided to examine rainfall data for ten randomly selected days in July and August in 2007 to make a choice. The amounts of rainfall (in millimeters) for the two months are shown below.

| Day | Rainfall in July ( mm) | Rainfall in August ( mm) |
|---|---|---|
| Week 1, Day 1 | 32 | 25 |
| Week 1, Day 3 | 35 | 31 |
| Week 2, Day 2 | 35 | 35 |
| Week 2, Day 4 | 37 | 37 |
| Week 2, Day 7 | 37 | 37 |
| Week 3, Day 2 | 37 | 37 |
| Week 3, Day 5 | 38 | 38 |
| Week 3, Day 7 | 39 | 39 |
| Week 4, Day 5 | 40 | 42 |
| Week 4, Day 6 | 40 | 49 |

i. Based on the information, which month should the Ministry choose, given that they would want a month that has a consistently low amount of rainfall?

ii. A few days later, the Ministry re-looked at the data and realized that they made a mistake for the figure recorded Week 4, Day 6 in July. Instead for 40 mm, the rainfall should be 60 mm. Given this new figure, which month should the Ministry choose now, if they want one that has a consistently low amount of rainfall?

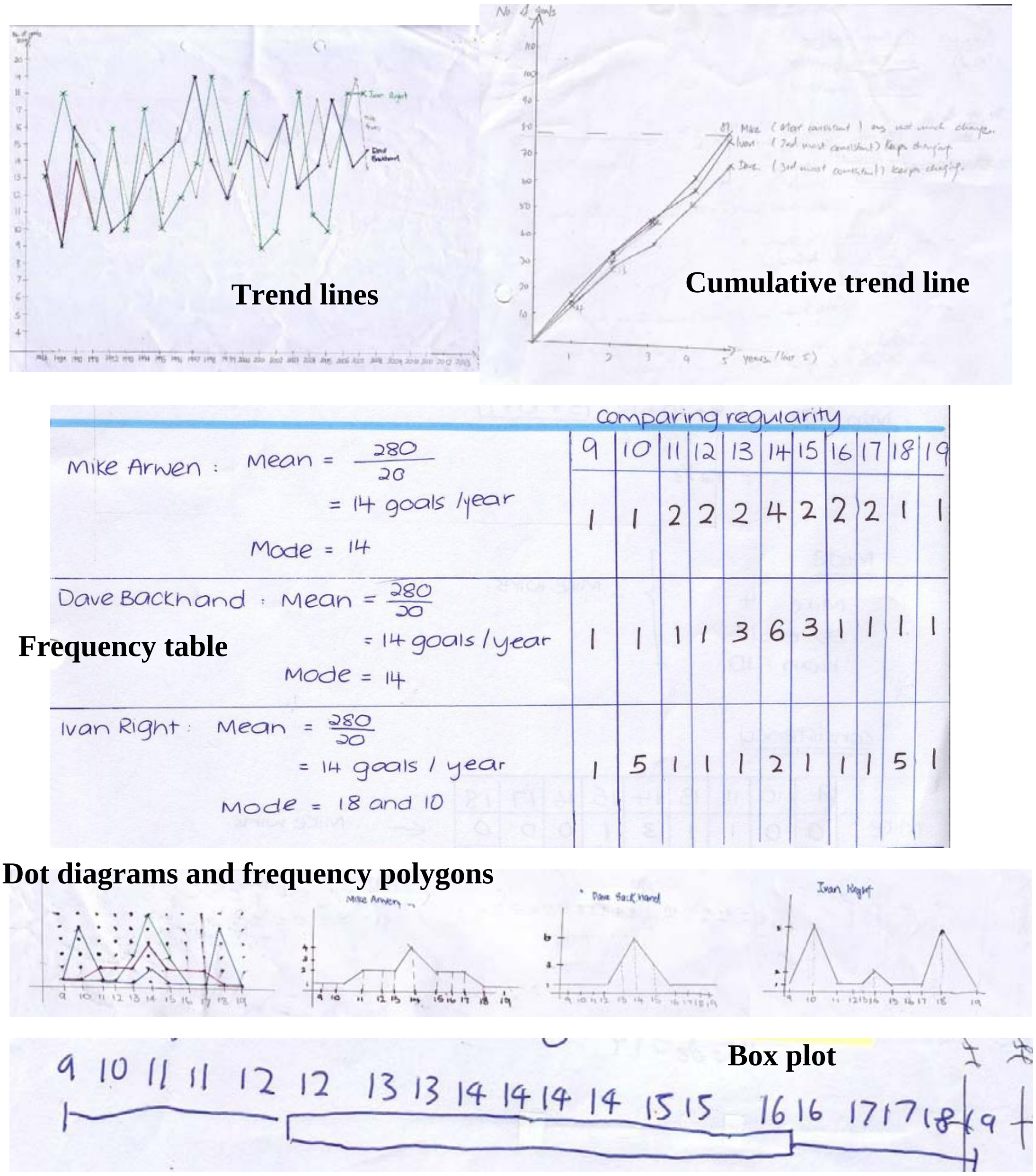
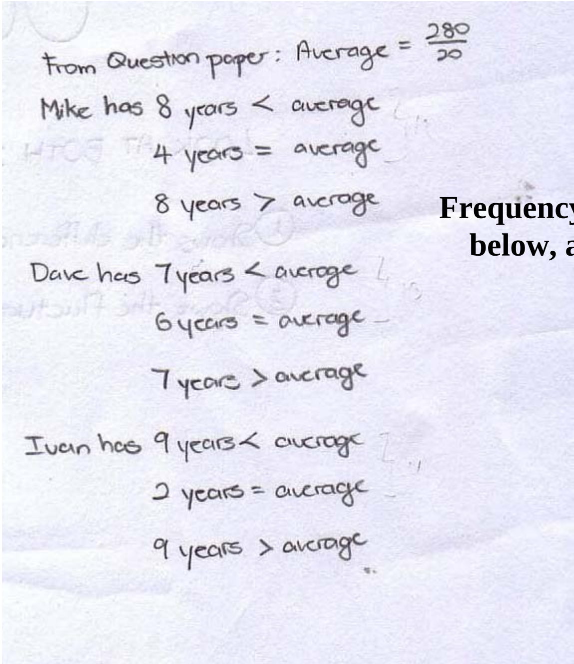
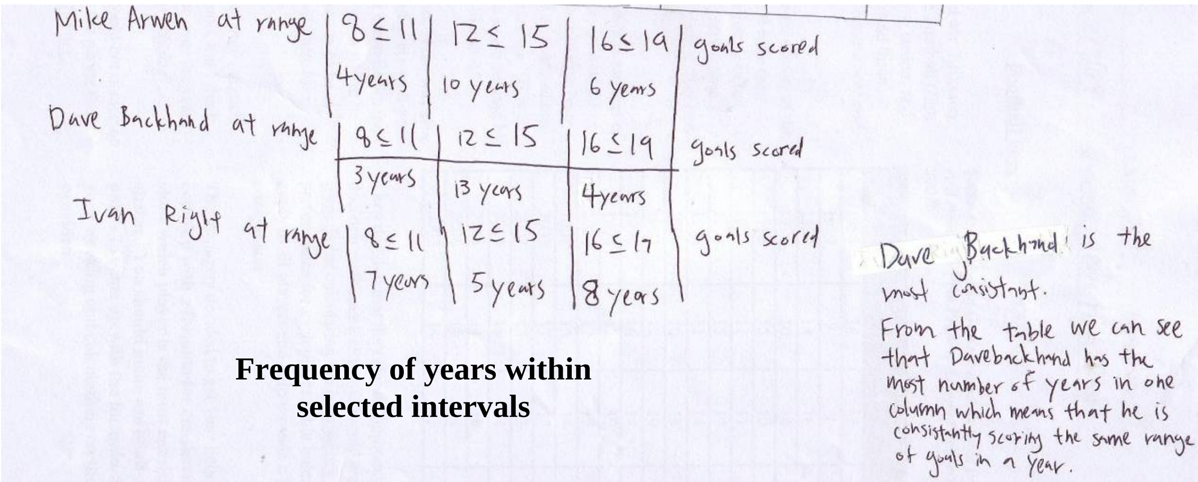



*Figure 1*. Examples of qualitative representations and methods for measure consistency. Students generated a diversity of representational forms such as trend lines, frequency tables, dot diagrams, frequency polygons, and box plots.




**Frequency of years above, below, and at average**




**Frequency of years within selected intervals**

*Figure 2*. Examples of frequency representations and methods for consistency measure

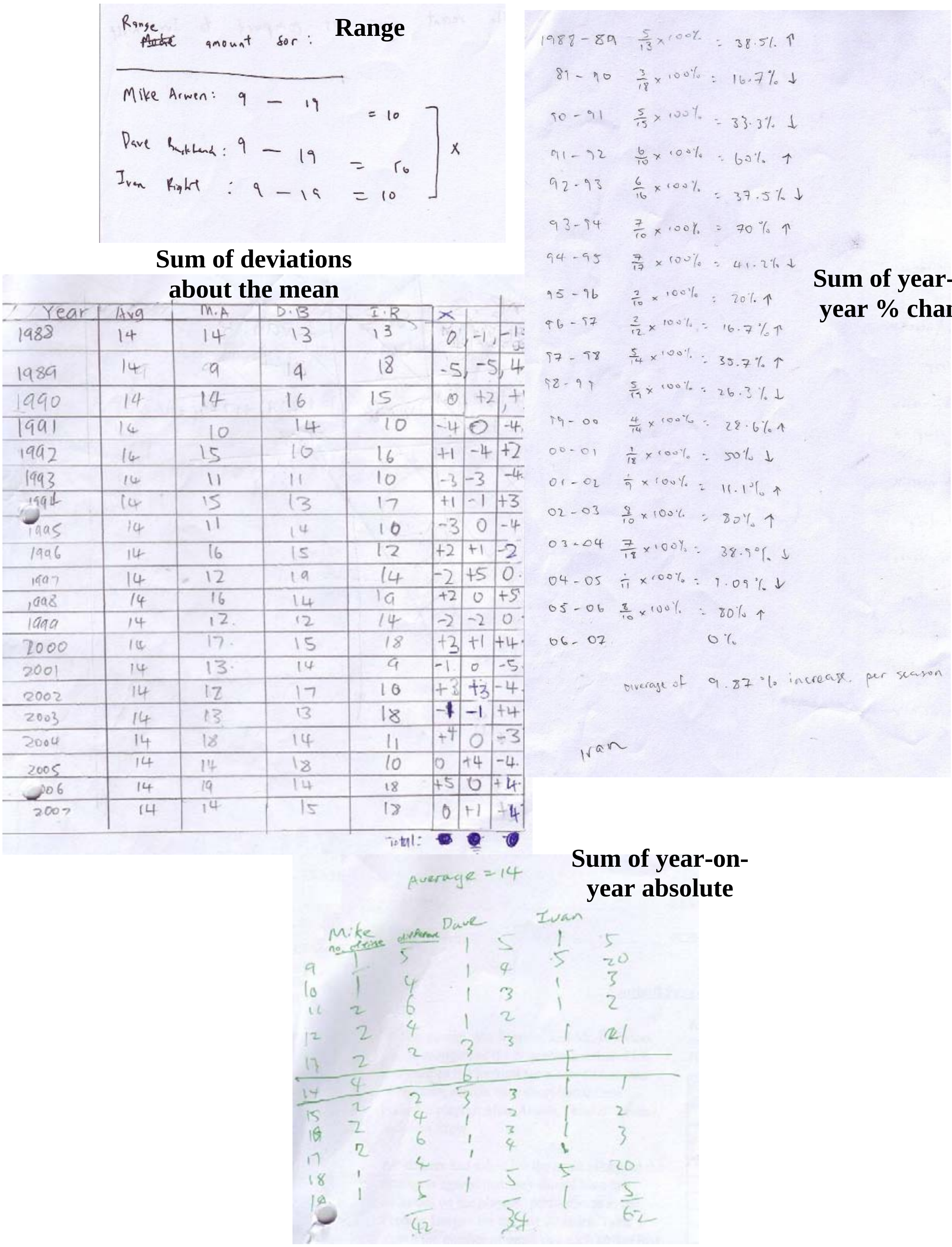



*Figure 3*. Examples of deviation-based representations and methods for consistency measure. Students generated a diversity of representational forms such as range, year-on-year percentage change, year-on-year deviation, and year-on-year absolute deviation.

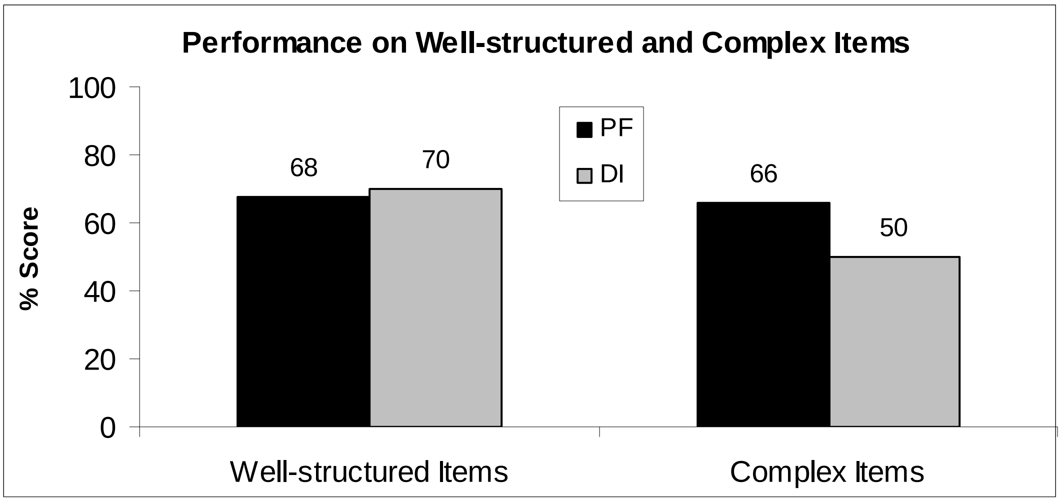


*Figure 4.* Breakdown of posttest performance as a *percentage* of the maximum score for the well-structured and complex, data analysis items