Journal of the Association for Information Systems

Volume 2, Article 4
July 2001

AN EXPERIMENT IN COLLABORATIVE DEVELOPMENT TO REDUCE SPREADSHEET ERRORS

Raymond R. Panko
University of Hawaii
panko@hawaii.edu

Richard P. Halverson, Jr.
Guide.Net, Inc.
rich@guide.net

Abstract

To study the extent to which group development can reduce spreadsheet errors, an experiment compared error rates in spreadsheet development by subjects working alone (monads) and by subjects working in groups of three (triads). Impressively, triads made 78% fewer errors than monads. However, this was not as large a reduction as nominal group analysis suggests was possible. Members of triads were satisfied with group development. However, triads whose work went most smoothly, whose members were most satisfied with group interactions, and that had the loosest leadership structure were significantly *more* likely to make errors than other triads.

Keywords: Spreadsheet, spreadsheet error, end user computing, human error, satisfaction, leadership

I. INTRODUCTION

We normally think of collaboration in terms of multi-user tools, such as as groupware and videoconferencing. However, collaboration may also be desirable for "personal productivity" applications, such as spreadsheet programs, because of its ability to bring multiple viewpoints to bear on tasks. In fact, an ethnographic study of spreadsheet developers (Nardi and Miller 1991) found that collaboration is common in development, both to obtain help for difficult parts of the development process and to have someone check output for reasonableness, in order to catch errors.

The last point raises the prospect that groupwork may be able to reduce errors in spreadsheet development. In recent years, there has been growing evidence that errors in spreadsheets are commonplace. Table 1, taken from the Spreadsheet Research (SSR) website (Panko 2001b) shows data from seven field audits that collectively examined 367 real-world spreadsheets. The Spreadsheet Research website also lists data from experiments in which almost 1,000 subjects built over 1,000 spreadsheets. Note that field audits found errors in 24% of the operational spreadsheets they examined, and the most recent field audits, which tend to use more effective auditing techniques than older audits, have found errors in 91% of the 54 spreadsheets they examined. Note also that the errors recorded were not trivial. Most audits limited their reporting to material errors.

Table 1. Field Audits of Spreadsheets

Study	Year	Spreadsheets	Pct w Errors	Cell Error Rate (CER)
Davies and Ikin[a]	1987	19	21%	NR
Cragg and King[b]	1993	20	25%	NR
Hicks[b]	1995	1	100%	1.2%
Butler[c]	1995	273	11%	NR
Coopers and Lybrand[d]	1997	23	91%	NR
KPMG[e]	1997	22	91%	NR
Lukasik	1998	2	100%	2.2%, 2.5%
Butler[b]	2000	7	86%	0.38%
Overall	NA	367	24%	
1997 and Later	NA	54	91%	

Notes: NR = Not reported.
[a]"Serious errors only"
[b]1.2% error in a multi-billion dollar spreadsheet
[c]Only reported errors large enough to demand additional tax payments
[d] Only reported spreadsheets off by at least 5%
[e] Only reported "major errors"

In four cases, the percentage of incorrect cells in audited spreadsheets was recorded. For three audits, the figure was one to three percent. In one audit, it was only 0.38%. These error rates are similar to error rates in programming and other nontrivial human cognitive activities (Panko 2001a). Error rates found in spreadsheet development experiments are also similar (Panko 2001b). These similarities should not be surprising. Human error theory (Reason 1990) has shown that the both correct performance and occasional errors are due to the same cognitive mechanisms. Error, in other words, is due to the fundamental ways we think, not merely to sloppiness. Although these error rates are not surprising in light of past results, they mean that nearly every large spreadsheet is likely to contain at least one material error and that even relatively small spreadsheets of a few dozen cells are likely to contain an error.

The implications of spreadsheet errors are sobering. Each year, tens of millions of spreadsheet users around the world create hundreds of millions of

spreadsheets. Although most of these spreadsheets are small throwaway calculations, quite a few are large (Cragg and King 1993; Gable et al. 1991; Hall 1996), complex (Hall 1996), and important to the entire organization, not just to the spreadsheet developer (Chan and Storey 1996; Gable et al. 1991; Hall 1996).

Can groupwork reduce spreadsheet errors? Steiner (1972) has shown that, for mathematical problem solving, groupwork should be able to reduce errors considerably. If *n* people working alone would each have a probability *e* of making an error for a given task, then if they work in a group, their probability of making an error should fall to e^n. To give a concrete example, suppose that each person working alone will have a 5% chance of making an error when entering a formula. With two people, the probability of an error should fall to only 0.25%. With three people, it should fall to a mere 0.0125%.

In programming, two experiments have already shown that when programmers work in pairs, their program quality increases (Nosek 1998; Williams 2000). In fact, "pair programming" (two-person development) is a major tenet of the *extreme programming* methodology (Beck 2000). In spreadsheeting, Panko and Halverson (1997) had subjects develop spreadsheets working alone (monads), in pairs (dyads), and in groups of four (tetrads). Tetradic work reduced errors substantially. However, as will be discussed later, the Panko and Halverson study was only a pilot study and had a number of experimental flaws. The study reported in this paper replicates the Panko and Halverson study with better experimental controls and compares people working alone (monads) to people working in groups of three (triads).

II. RESEARCH ISSUES

RESEARCH MODEL

Figure 1 illustrates our preliminary research model. It shows that we expected group development to influence two dependent variables: adoption and error rates. We will now look at the elements of this model in more detail.

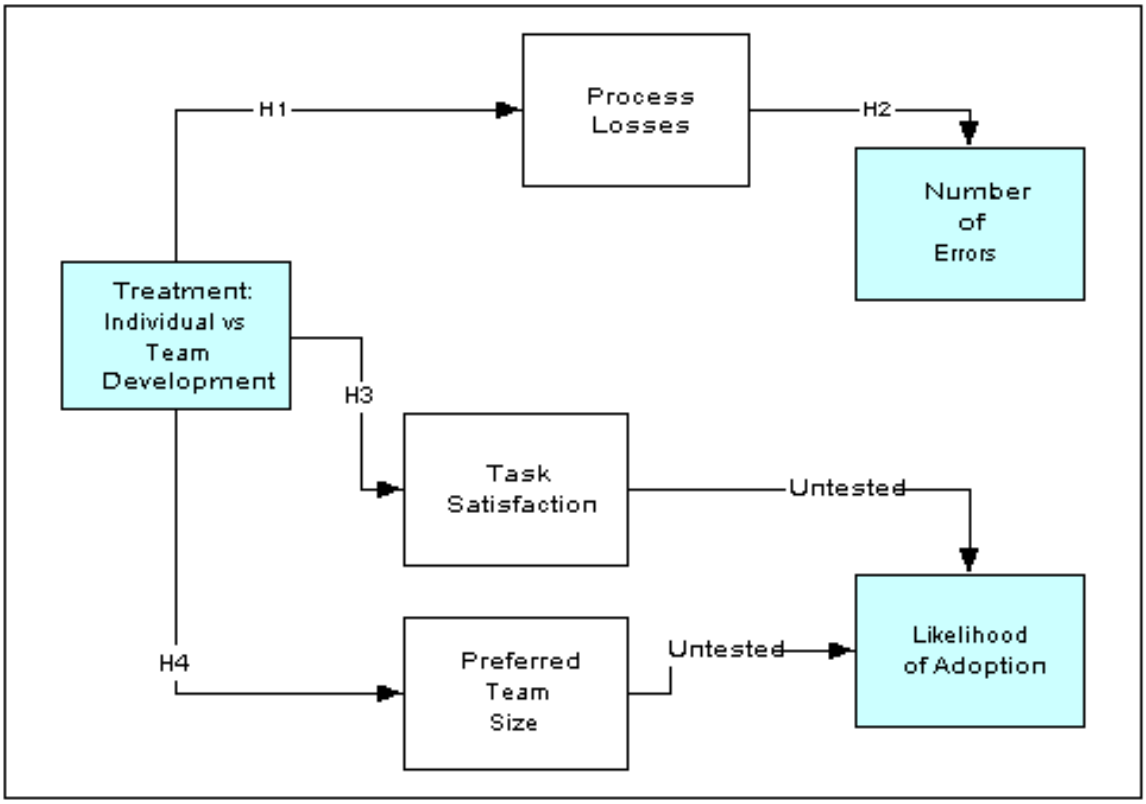

Figure 1. Initial Research Model

DEPENDENT VARIABLES

We will begin at the left of the picture, looking at the dependent variables.

Group Error Reduction

Following the Steiner's (1972) argument discussed earlier, we expect that groups will make fewer errors in spreadsheet development than will individuals working alone. This leads to our first hypothesis.

H1: Groups will make fewer spreadsheet development errors than will individuals working alone.

Process Losses and Nominal Groups

However, teams do not always achieve their potential. Steiner (1972) referred to the measured gap between theoretical group performance and actual performance as the group's *process losses*. The traditional way to measure process losses is to use nominal groups (Marquart 1955). With this technique, some subjects work in actual groups. Their performance gives us a measure of actual performance.

Other subjects work alone. Then, if actual groups are of size *n*, the data from *n* subjects working alone are combined into those of a *nominal group*, that is, a group in name only. The results of the individuals working alone are combined in a simple way. If *any* of the nominal group members working alone did the task correctly, then the nominal group should have the resources needed to do the task correctly, so the nominal group is credited as doing that task correctly. This gives us a measure of theoretical group performance. Process losses should make actual group performance lower than nominal group performance, leading to Hypothesis 2.

H2: Nominal groups will make fewer errors than actual groups.

Adoption

Even if group spreadsheet development could reduce errors, this potential would be valueless if developers refused to engage in group spreadsheet development or if they resisted strongly. Certainly, group spreadsheet development is a new way to develop spreadsheets, so it is important to know how people would react to it.

INTERMEDIATE VARIABLES

The arrows leading to likelihood of adoption are marked "untested" to indicate that they were not studied in this experiment. However, it seems reasonable to assume that adoption will be influenced by two intermediate variables: satisfaction and preferred group size.

Satisfaction with the Experience

It is plausible that if people are pleased with the experience of group development, they are likely to adopt it. Therefore, it seemed important to have satisfaction as an intermediate variable leading to likelihood of adoption.

Preferred Group Size

A related assumption is that if people who experience group development prefer groupwork after the experience, they are more likely to adopt in group development.

GROUP DEVELOPMENT EXPERIENCE, SATISFACTION, AND PREFERRED GROUP SIZE

This study focuses on the implications of group development. We have already seen that our model assumes that group development will decrease the number of errors compared to individual development. The model in Figure 1 shows that we expect that the experience of group development will tend to increase satisfaction and make people more likely to believe that group development is better than individual development.

A survey of research comparing group size with satisfaction (Wheelan and McKeage 1993) has shown consistent findings that satisfaction falls with group size. However, studies in this survey did not look at groups of size one, that is, people working alone, when comparing satisfaction levels.

In contrast, studies of brainstorming (Stroebe et al. 1992) have shown that people believe that they are more productive working in brainstorming groups than when working alone (despite strong evidence that group brainstorming is less effective than brainstorming with nominal groups) and prefer to work in live brainstorming groups.

In addition, two experiments in pair programming (Nosek 1998; Williams 2000) and the Panko and Halverson (1997) spreadsheet development experiment

also found that team programmers preferred group development to solitary development.

Because only brainstorming and multi-person programming studies compared individuals to groups and found greater satisfaction in groups, we created Hypotheses 3 and 4.

H3: Subjects developing spreadsheets in groups will have higher overall satisfaction than will subjects working alone.

H4: Subjects will prefer working in groups to working alone.

EXPLORATORY ANALYSES

In addition to hypothesis testing, we also engaged in some exploratory research, denoted by the letter "e" in Figure 1. We did not have formal hypotheses for this exploratory research, although we had informal expectations.

Satisfaction, Group Interactions, Process Difficulties, and Leadership

Although, as discussed later, we had a single question that measured overall satisfaction for purposes of testing Hypothesis 3, satisfaction is a complex concept. We asked a number of questions related to satisfaction with group interactions, for instance, asking the subject whether he or she agreed with the statement, "The group was accepting and nonjudgmental."

We were also concerned with process difficulties if people worked in groups. Accordingly, we included a number of process difficulty questions, for instance, "We had trouble pointing to things on the screen."

Another potential influencer of satisfaction is the presence of leadership. Leadership might improve satisfaction by creating an orderly environment. On the other hand, leadership might prove to be constraining and, therefore, reduce satisfaction.

Performance, Group Interactions, Process Difficulties, and Leadership

Although the initial model in Figure 1 does not indicate an assumed correlation, it is possible that performance (number of errors) could be influenced by group interactions, process difficulties, and the presence of leadership. In a very exploratory analysis, we correlated questions in these areas with whether the group got the correct answer or was wrong.

III. METHODOLOGY

BACKGROUND: THE PANKO AND HALVERSON PILOT STUDY

As noted earlier, our methodology builds on an earlier study by Panko and Halverson (1997), who had subjects build spreadsheets from a common task statement, which they called "Galumpke." This task required subjects to build a two-year pro forma corporate income statement. Subjects worked alone (monads), in groups of two (dyads), or in groups of four (tetrads). Dyads made only 32% fewer errors than individuals, and this modest difference was not statistically significant. However, tetrads made 65% fewer errors than people working alone. This difference was statistically significant, and its size was of practical importance.

Although the Galumpke study results were interesting, it was only a pilot study. Most importantly, it sacrificed several experimental controls. For instance, the study allowed some subjects to do their work outside the laboratory, raising the possibility of cheating. In addition, most subjects who worked alone used Microsoft Excel, while all tetrads used Lotus 1-2-3.

The wording of the Galumpke task may also have caused problems. This is how the task was worded:

> Your task will be to build a two-year pro forma income statement for a company. The company sells galumpkes, which are small food warmers used in restaurants. The owner will draw a salary of $80,000 per year. There is also a manager of operations, who will draw a salary of $60,000 per year. The income tax rate is expected to be 25% in each of the two years. Each galumpke will require $40

> in materials costs and $25 in labor costs in the first year. These numbers are expected to change to $35 and $29 in the second year. There will be a capital purchase of $500,000 in the first year. For depreciation, assume 10% straight-line depreciation with no scrap value. Unit sales price is expected to be $200 in the first year and $180 in the second year. There will be three sales people. Their salary per person is expected to average $30,000 in the first year and $31,000 in the second. The rent will be $3,000 per month. The company expects to sell 3,000 galumpkes in the first year. In the second, it expects to sell 3,200.

One problem was that only half the subjects in the Galumpke study knew to handle one part of the task, namely the treatment of capital purchases and depreciation. In addition, mistreating the capital purchase often resulted in negative income, creating additional problems for the treatment of income tax. Confusingly, some subjects applied the income tax rate to salaries rather than to corporate income, perhaps because the task wording referred to "income tax" rather than "corporate income tax." Finally, the wording describing the salaries for the firm's three salespeople may have been misleading, leading some subjects to treat the per-person sales worker salary as the total salary of three sales workers.

The experiment reported in this paper builds on the Galumpke pilot study. Given the poor performance of dyads in the Panko and Halverson experiment, and given the high cost of tetradic work, the current study had subjects work alone (monads) or in groups of three (triads). In addition, the current study had all subjects work in the laboratory and use Excel.

THE MICROSLO TASK

The current study revised the Galumpke task to eliminate wording problems noted above. We call the revised task "MicroSlo" to distinguish it from the Galumpke task. Here is the wording for the MicroSlo task:

> Your task is to build a two-year pro forma income statement for a company. The company sells microwave slow cookers, for use in restaurants. The owner will draw a salary of $80,000 per year. There

is also a manager of operations, who will draw a salary of $60,000 per year. The corporate income tax rate is expected to be 25% in each of the two years. Each MicroSlo cooker will require $40 in materials costs and $25 in labor costs in the first year. These numbers are expected to change to $35 and $29 in the second year. Unit sales price is expected to be $200 in the first year and to grow by 10% in the second year. There will be three sales people. Their salary is expected to average $30,000 per person in the first year and $31,000 in the second. Factory rent will be $3,000 per month. The company expects to sell 3,000 MicroSlo cookers in the first year. In the second, it expects to sell 3,200.

THE SAMPLE

Our sample consisted of undergraduate business students in four sections of a second course in end user computer skills. All were first-semester students in the College of Business Administration of a state university. All were at least third-year university students. All had taken two accounting courses and a computer skills course that taught spreadsheet development. In addition, in their current course, they had already completed a refresher module on spreadsheet development.

Students were required as part of their class grade to participate in an experiment or complete an alternative assignment. Of 143 students, 103 participated in the experiment. No demographic, accounting, or spreadsheet experience differences were found between students who chose to participate in the experiment and those who did not. One spreadsheet created by an individual was excluded because it contained no formulas. Following the practice in the Galumpke study, 22 accounting and finance majors were excluded because of their specialized knowledge of pro forma income statements. The remaining 80 students were called "general business students."

THE PROCEDURE

When subjects arrived, the purpose of the experiment was explained. After being invited to ask questions about the experiment, subjects filled out an

agreement to participate and a preliminary questionnaire. General business students were randomly assigned to one of the two conditions. Subjects working alone (monads) worked in an electronic meeting room. They could not see the spreadsheets being developed by other subjects. No conversation was permitted in the room during the experiment. Triads worked in breakout rooms. The general business students working alone produced 35 useable spreadsheets. The 45 general business students working in triads produced 15 spreadsheets.

For maximum statistical power, one should assign equal numbers to each condition. However, for some hypotheses (H1 and H2), there should have been equal numbers of *spreadsheets,* while for other hypotheses (H3 and H4) there should have been equal numbers of *subjects.* The actual ratio of students working alone and in triads was a compromise between the optimum 3:1 and 1:1 ratios for these two types of comparisons.

Subjects were given the task and told that they would have to build spreadsheets from the task statement. They had previously completed a word problem task as a homework assignment. They were also told that they would have to design the spreadsheets on their own, without help. They were urged to work as carefully as possible but would receive full credit even if they made errors, so long as they did their best to work as accurately as possible. All students used Excel, which they had used in previous course homework.

When subjects finished the MicroSlo task, they were given a post-experiment questionnaire that asked about their experience. With a few exceptions that will be noted later, all questions used seven-point Likert scales, with 7 being the highest value and 1 being the lowest value.

In the post-experiment questionnaire, the triad members reported that they had sufficient time (mean 6.3 out of 7). Subjects working alone had the same mean. Triad members disagreed that they had a difficult time using Excel (1.8), and so did subjects working alone (2.1). For the questions asking if the members of the triad

knew one another before the experiment, the mean was only 1.4, indicating that most teams consisted of strangers or near strangers.

ERROR RECORDING

To analyze the spreadsheets, the first author opened each and checked the bottom-line values against the standard solution. Errors were fixed and recorded until the bottom-line figures were correct. Because it is unlikely that any combination of errors could produce the correct bottom line, this method probably caught almost all errors. Hypothesis testing used the one-tailed Excel t-test function with unequal variances. (Note to reviewers. We also did the analysis with comparable nonparametric test [Mann-Whitney-Wilcoxon] and the results were identical to three decimal points. It seems best to report the t values because readers will be more familiar with them.)

CELL ERROR RATES (CER)

As our error rate measure, we use the **cell error rate (CER)**, which is the percentage of cells containing errors. Only the first occurrence of each error is counted. The CER gives us an error rate measure comparable to the *faults per thousand lines of code (faults/KLOC)* metric commonly used to discuss error rates in software. Historical averages for faults/KLOC, multiplied by the program size, can be used to estimate roughly the number of faults that can be expected in a program under development. Although faults/KLOC forecasts need to be modified for program complexity and second-order size effects to be fairly accurate (Ferdinand 1993), even an unmodified forecast gives a good rough estimate of program faults. Similarly, CER should give us a way to compare error rates in spreadsheets of different sizes.

Our reporting of cell error rates (CERs) is slightly different from that in the Galumpke (Panko and Halverson 1997) pilot. That study divided the number of errors by the actual number of cells in the spreadsheet. In this study, we divided the

number of errors by the size of a model solution, which has 36 cells. Using model solution size in the denominator makes comparisons easier across studies that use the same task.

IV. RESULTS

HYPOTHESIS TESTS

Table 2 shows the results of the current study. It also shows the results of the Galumpke (Panko and Halverson 1997) pilot study. The Galumpke study used a slightly different task and, as just noted, had a slightly different basis for computing CERs. However, as the table shows, the current study repeats the general trends seen in the earlier study.

Table 2. Patterns of Errors

	This Study			Panko and Halverson (1997)		
	General Monads	**General Triads**	**Nominal Triads**	**General Monads**	**General Dyads**	**General Tetrads**
Subjects	35	45	33	42	46	44
Spreadsheets	35	15	11	42	23	11
Spreadsheets with errors (percent)	86%	27%	0%	79%	78%	64%
Errors per spreadsheet (mean)	1.8	0.4	0.0	2.36	1.61	0.82
Cell error rate (mean)	4.6%	1.0%	0.0%	5.6%	3.8%	1.9%
CER Improvement vs. general alone	NA	78%	100%	NA	32%	66%

Notes: The Galumpke pilot (Panko and Halverson 1947) used a different version of the task used in this study.
It also used a slightly different way of calculating the cell error rate (CER).

Terms: General monad: Subjects worked alone. Accounting limited to preparatory courses.
General dyad: Subjects worked in groups of two. Accounting limited to preparatory courses.
General triad: Subjects worked in groups of three. Accounting limited to preparatory courses.
General tetrad: Subjects worked in groups of four. Accounting limited to preparatory courses.
Nominal triad: Data aggregated from three general subjects working alone.

Probabilities for t-tests based on number of errors: 0.00001 General alone versus general triads
0.027 General triads versus nominal triads

Individuals versus Triads

Table 2 shows that triads did substantially better than general business students working alone (monads). Triads had errors in only 27% of their spreadsheets instead of 86% for the monads and had a cell error rate (CER) of 1.0% instead of 4.6% for monads. The improvement in the cell error rate was 78%.

The difference in errors per spreadsheet between triads and monads was statistically significant ($t = 4.45$, $df = 22$, $p = 0.0001$). Therefore, the null hypothesis was rejected and hypothesis HI, that subjects working in groups make fewer errors than when they work alone, was accepted. In addition to being statistically significant, the difference is large enough to be practically important.

Actual versus Nominal Groups

Still, it appears that triads could have done even better. Eleven nominal groups were constructed from 33 monad spreadsheets using random selection without replacement. Among these nominal group spreadsheets, *none* had errors in any subtask based on the nominal group methodology described earlier. In other words, at least one of the three members of each nominal group had the correct answer for each subtask.

Although there were only 11 nominal groups and 15 real groups, the difference in number of errors per spreadsheet between nominal triads and actual triads was significant ($t = 2.56$, $df = 14$, $p = 0.021$). So for our second hypothesis, the analysis rejected the null hypothesis and accepted H2, that nominal groups make fewer errors than real groups. This is disappointing, because it means that it should have been possible to reduce errors by 100% instead of the 78% that our actual triads reduced errors.

In the Galumpke study, Panko and Halverson observed many groups developing their spreadsheets. They noted that when the typist made a thinking error, typed the wrong number, or pointed to the wrong cell, the other team members often were looking away to engage in a side discussion, reading the task

sheet, or simply looking away for no apparent purpose. Errors happened so rapidly that they were often undetected by other team members. The authors of this study also observed the triads working, and it was clear that subjects in our triads often looked away from the screen.

SATISFACTION WITH GROUPWORK

Because group development would be difficult to implement if people resisted it strongly, this study asked subjects a number of questions about their satisfaction. One was a general question asking the subject to rate how satisfied they were with the task. This was a seven-point Likert scale, as noted above. Subjects working in triads had a slightly higher mean satisfaction level (5.2) than subjects working alone (4.8). However, this difference was not quite statistically significant ($t = 1.25$, $df = 70$, $p = 0.108$), although it was in the expected direction. Because the result did not reach significance, we rejected H3, that subjects in groups have higher overall satisfaction than subjects working alone.

At a practical level, however, the fact that members were generally happy with group work was encouraging, given the fact that past research, as noted earlier, generally has shown *reductions* in satisfaction with group size.

Our final hypothesis was that subjects who worked in groups would prefer groupwork to working alone. We asked subjects to estimate the best group size to do the task. Among the subjects who worked in tetrads, the mean preferred group size was 2.5. This mean was statistically higher than one ($t = 13.69$, $df = 42$, $p < 0.0000$), which would represent working alone. For H4, we rejected the null hypothesis and accepted the alternative hypothesis that subjects who experienced groupwork would prefer group work to working alone. In fact, only five of the 45 subjects who worked in triads said that a size of one would be best for this task. Even among subjects working alone, the average preferred group size was 1.7, and only a third of the monads thought that working alone was the best group size for the task.

In terms of the acceptability of group spreadsheet development, then, subjects who actually worked in triads were comfortable with group development, and even subjects working alone seemed to think it would be a good idea.

REACTIONS TO THE EXPERIMENT

Although the main focus of the study was hypothesis testing, we asked some general questions about the reactions of subjects who worked in triads to the group spreadsheet development experience. We discuss their responses in three categories: satisfaction with groupwork, leadership, and process difficulties. As noted earlier, all responses were based on Likert scales with 1 as the low value and 7 as the high value.

Satisfaction with Groupwork

Overall, subjects working in triads were happy with their groupwork experience. They felt fairly strongly that their groups were relaxed and not uptight (mean = 6.2 out of 7) and rated their overall satisfaction with group interactions very highly (6.1). They generally were satisfied with the group's performance (5.8) and felt that their group was pleasant and nonjudgmental (5.9). They reported fairly little hostility (2.5) or competition (2.8), and they disagreed that they worked as individuals rather than as groups (2.7). There were some feelings of being inhibited from expressing feelings (3.5), but we did not ask whether they felt that this inhibition was personal or group-driven.

Leadership

The leadership results are interesting because they indicate both that group leaders emerged (mean = 4.8) and that there was no specific leader (4.5). However, these two statements had a strong *negative* correlation (-.617), indicating that the presence of a leader varied considerably across groups.

Process Problems

The only common process problem was difficulty in coming to decisions. However, this is not surprising given the difficulty of the task, and the means indicate that problems in this area were fairly mild (1.8 to 2.8) for several questions that probed this domain.

We were concerned that sharing a single computer would cause problems, but there were only modest complaints about difficulties in sharing the computer (2.0) or in pointing to things on the screen (1.8). However, there was strong agreement with the statement that one person did most of the typing (6.1), and observation of the groups confirmed that groups shifted typists only rarely. In other words, groups largely ignored the possibility of shifting control of the computer. In almost all cases, one member who felt confident about his or her ability to do the task took the keyboard initially. If they kept their own confidence and the group's confidence, they kept the keyboard. However, in four groups, the person initially controlling the computer lost their own confidence or the group's confidence, and someone else took over. The turn-over was smooth and rancor-free in all four groups. It was simply an obvious thing to do based on the recognition of expertise by group members.

CORRELATES OF ERRORS

Going beyond our research hypotheses, we correlated a number of variables against the number of errors made by subjects. We did so only for triad data. We had not expected to see any correlations because only four of the 15 groups made even a single error, and only six errors were made in total.

However, we were surprised by the results in both effect sizes and more fundamentally, in effect directions. Table 3 shows the most strongly significant correlations between post-experiment questionnaire variables and the number of errors made by the group. All variables shown had an uninflated cutoff of 0.10 or better. This relatively high cut-off was used instead of the customary 0.05 because

of the exploratory nature of the analysis. Actual probabilities are listed if the reader prefers a stricter cut-off criterion. Fourteen variables had Pearson correlations that were significant at this level. We correlated the number of errors in the spreadsheet against 79 variables, so at least some of the 14 "significant" correlations may be spurious.

Table 3. Significant Correlates of Number of Errors per Spreadsheet

	Mean	*Correlation w Number of Errors*	*t value*	*tails*	*P Errors*
Wrong	0.29	0.921	13.62	1.00	0.000
% other triads w errors	21.36	0.438	2.80	1.00	0.004
I was interested in the task	4.76	-0.399	2.50	1.00	0.009
Tendency to talk at same time	2.29	**-0.426**	2.71	2.00	0.011
% wrong if individuals	33.18	0.339	2.07	1.00	0.023
Group's knowledge of Excel	5.78	-0.336	2.05	1.00	0.024
I was group's leader	2.64	-0.348	2.14	2.00	0.040
Trouble pointing to things on screen	1.78	**-0.345**	2.11	2.00	0.042
Much disagreement	2.02	**-0.333**	2.03	2.00	0.050
After built, checked all cells for errors	3.93	-0.273	1.63	1.00	0.056
Difficult time using Excel	1.80	**-0.271**	1.62	1.00	0.058
G accepting and not judgmental	5.90	**0.306**	1.85	2.00	0.074
Group's knowledge of accounting	5.11	-0.246	1.46	1.00	0.077
I felt tense and uncomfortable	2.42	**-0.298**	1.79	2.00	0.082
Felt I could express disagreement freely	5.69	**-0.296**	1.78	2.00	0.084

Some of the variables that correlated with the number of errors were unsurprising. For instance, errors were correlated positively with the subject's estimate of the percentage of *other* triads ($r = .438$, $p < 0.001$) and monads ($r = .339$, $p = 0.023$) that had created incorrect spreadsheets (but interestingly not with the subject's estimate of the probability that *his or her own triad* had made an

error). This is consistent with general findings that most people regard themselves as above average across a wide variety of activities (Brown 1990).

In other non-surprising correlations, errors fell with greater interest in the task ($r = –.399$, $p = 0.004$), with the subject's assessment of the group's knowledge of Excel ($r = –0.336$, $p = 0.024$), and with checking the spreadsheet for errors after development ($r = –0.273$, $p = 0.056$)—although in fact no group made more than a cursory check after development—and with the group's knowledge of accounting ($r = –0.246$, $p = 0.077$).

More surprising was a group of variables related to satisfaction. We had no hypotheses about the relationship between satisfaction and the number of errors, but several satisfaction variables related to the number of errors, and they indicated that satisfaction generally is *positively* related to the number of errors: more satisfied people were, the *more* errors they made. Agreeing that there was much disagreement had a strong negative correlation with the number of errors ($r = –.333$, $p = 0.05$). In contrast, agreeing that the group was accepting and nonjudgmental correlated positively with the number of errors ($r = 0.306$, $p = 0.074$). Finally, the subject's belief that they could express disagreement freely was negatively correlated with the number of errors ($r = –0.296$, $p = 0.084$).

Another interesting trend in Table 3 is a negative correlation between the number of errors and having difficulty. In general, the more difficulty the group had, the fewer errors it made. For example, a tendency for people to talk at the same time was negatively correlated with the number of errors ($r = –.426$, $p = 0.011$). Having problems pointing to things on the screen was also negatively correlated with the number of errors made by the triad ($r = –0.345$, $p = 0.042$). Feeling tense and uncomfortable was likewise negatively correlated with the number of errors ($r = –0.298$, $p = 0.082$). Even having a difficult time using Excel was negatively correlated with the number of errors made by the group ($r = –271$, $p = 0.058$).

There seemed to be three distinct clusters of other variables, although the sample size was too small for multivariate analysis. First, errors *fell* with what

appeared to be difficulties in process coordination. The more problems there were, the *fewer* errors the group made. These problems included the tendency for several people to talk at the same time (–0.426), trouble pointing to things on the screen (–0.345), and some members of the group talking too much (–0.255). These variables seem to be symptoms of a fully engaged group with bustling activity and people occasionally tripping over one another. For example, difficulty in pointing to things on the screen indicates that subjects were actively identifying errors or points of confusion. This seems to run counter to Putnam's (1986) suggestion that process conflict is bad, but Putnam was probably referring to process conflict so great that it paralyzed group progress.

Another group of variables indicates that the presence of a leader reduced errors . Several variables probed this dimension, for instance, the subject agreeing that they were the group's leader (–0.348), agreeing that a group leader emerged (–0.251), and agreeing that there was a definite leader (–0.251).

However, the group leadership variable is difficult to interpret. Observations of the groups suggested to the first author that a leader tended to emerge if someone in the group was highly knowledgeable. The group then accepted that person's leadership. Therefore, is the negative correlation between group leadership and errors due to the benefits of leadership per se or simply to the fact that someone in the group was very knowledgeable? Our data did not allow these competing interpretations to be resolved.

Third, there was a group of variables that related to interpersonal relations, but the directions of effects tended to be surprising. The presence of more interpersonal conflict *reduced* errors, for instance, the presence of much disagreement (–0.333), feeling tense and uncomfortable (–0.298), and feeling that the subject could express disagreement freely (–0.296). Errors *increased* if the subject said that the group was accepting and nonjudgmental (0.306).

The survey had one overall question on satisfaction with the group interactions. This variable's correlation with the number of errors was not quite

significant (r = .224, t = 1.50, p = 0.071), but it was positive, indicating again that greater satisfaction can lead to *more* errors.

Overall, the correlations shown in Table 3 indicate that error control was best in fairly active groups that had both discipline and conflict and in which relational satisfaction was sacrificed somewhat for performance. Again, however, this was an exploratory analysis.

V. POSSIBLE LIMITATIONS

We note two potential limitations in our study. One is that subjects worked in the laboratory, which is an artificial environment. However, for subjects who worked at home, the Galumpke pilot (Panko and Halverson 1997) got results that were very close to those in this study. In addition, field audits of operational spreadsheets built by real developers have always found substantial error rates (see Table 1).

The second possible limitation is that our subjects were undergraduate students with little or no spreadsheet experience at work. However results have been very similar from studies that used experienced developers (Brown and Gould 1987; Hicks 1995; Panko and Sprague 1998), studies that used students that had completed a course that taught spreadsheet development (Panko and Halverson 1997), and studies that used rank novices (Hassinen 1988, 1995). In addition, a code inspection study (Galletta et al. 1993) found that experienced spreadsheet developers were no more successful at finding errors than subjects with little or no spreadsheet development experience. One spreadsheet development experiment (Panko and Sprague 1998) directly compared undergraduates, MBA students with little or no spreadsheet development experience at work, and MBA students with considerable experience at work. It found no significant difference in error rate across these three groups. In general, human error research has shown that being an expert does not always reduce error rates significantly compared to being a novice (Panko 2001a; Shanteau and Phelps 1977; Wagenaar and Keren 1986).

VI. CONCLUSIONS

This study examined the impact of group spreadsheet development on two basic variables. The first was the number of errors made by groups. The second was the impact of groupwork on satisfaction, which arguably would affect the likely adoption of groupwork. Figure 1 presented our research model.

The model's expectation that group development would reduce errors was borne out clearly. When general business students worked in groups of three (triads) instead of working alone (monads), the percentage of incorrect spreadsheets fell 67%, from 86% to 27%. The cell error rate (CER)—the percentage of cells containing errors—fell 78%, from 4.6% to 1.0%.

The model's expectation that process losses would limit error reduction was also borne out. Nominal groups, which give us a measure of theoretically possible group performance, did significantly better than the actual triads, indicating that our triads were not getting the full benefit of groupwork. However, despite these losses, error reductions were still impressive.

Regarding likelihood of adoption, subjects who worked in triads did tend to believe that groupwork was better than individual work for this task. This acceptance of groupwork was quite strong.

The model, however, did not appear to be correct regarding satisfaction. Speaking narrowly, although triad members were slightly more satisfied than subjects working alone, the difference was not quite statistically significant.

More fundamentally, satisfaction appears to influence not only the acceptability of group work but also group performance. An exploratory analysis found that several satisfaction variables were correlated with group performance and tended to be *negatively* correlated with performance. In other words, greater satisfaction tended to be associated with more errors.

This suggests a revised research model (Figure 2) for future work. It shows a likely connection between satisfaction and both acceptance and performance. It

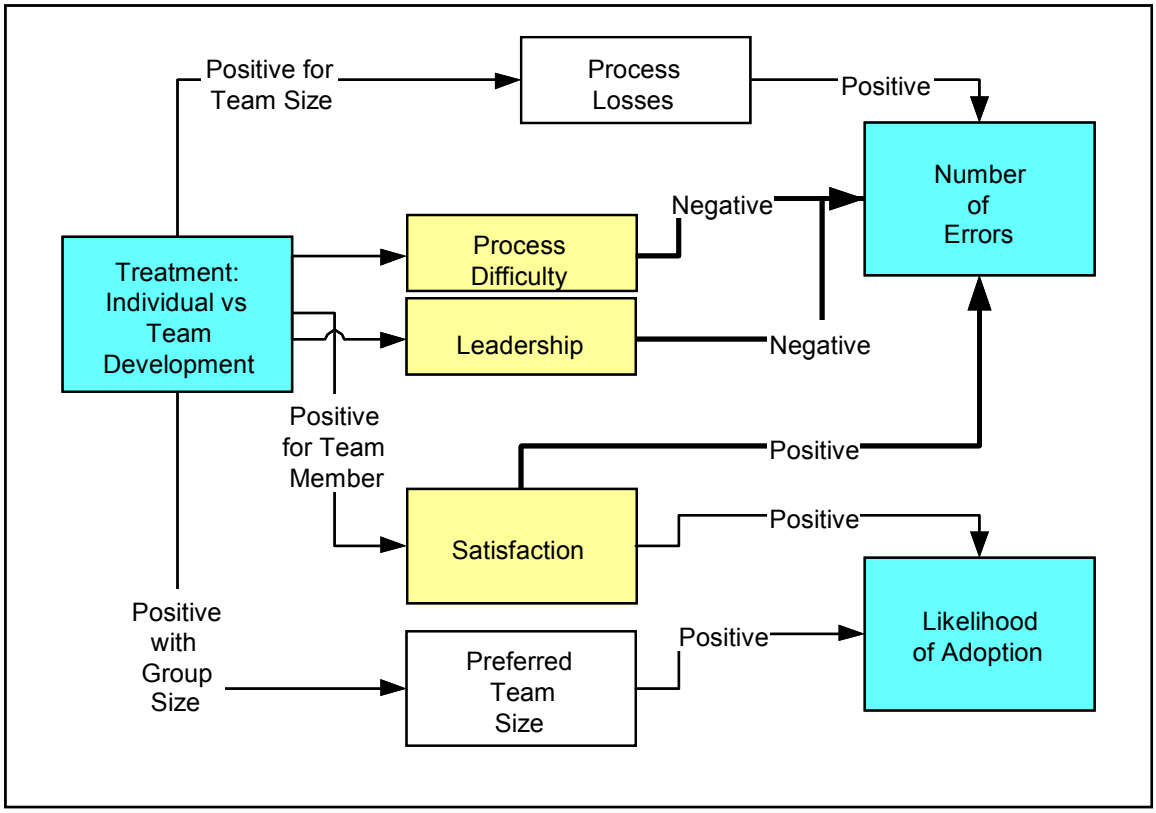

Figure 2. Revised Research Model

indicates that satisfaction may be a two-edged sword, making groupwork more acceptable but also tending to limit groupwork's error-reduction abilities.

However, satisfaction will need to be explicated more completely. There are many types of satisfaction, such as satisfaction with outcomes, satisfaction with process, and satisfacion with group interactions.

What did our subjects think about group development? Most of our subjects indicated that group development was preferable to individual development for this task. In addition, subjects who worked in triads generally were satisfied with group development. Although the idea of group spreadsheet development was novel to the subjects, they generally found it agreeable.

Still, our groups did not seem to be getting all of the benefits possible in groupwork. Nominal group analysis suggested that triads should have made no errors at all. There seemed to be significant process losses occurring. Why did our groups have process losses? One problem was that not every team member watched the screen at all moments. Because errors can happen very rapidly, even brief losses of vigilance can make groupwork less effective at detecting errors.

Another problem appears to be satisfaction leading to complacency and an unwillingness to challenge the mistakes of others. Although only four groups made any errors at all, there were significant correlations between the number of errors per spreadsheet and several satisfaction, leadership, and coordination variables. The groups that produced no errors had a disciplined yet still edgy environment, in which group members challenged one another and ran into some mild process difficulties because of their activeness.

Although the results are encouraging, one issue that remains is the size of group that will be best in spreadsheet development. Obviously there is a price-performance trade-off. Single-person development is the least expensive approach but has a high error rate. Adding developers reduces errors but also increases costs. The current study found that groups of size three were farily effective at reducing errors. The earlier pilot study by Panko and Halverson (1997) in contrast, suggested that groups of size two were not very effective. Unfortunately, although both the current study and the Panko and Halverson pilot study looked at groups ranging from one to four in size, the two studies used different tasks and methodologies and so are not directly comparable. A study that directly compares performance for development groups of different sizes is needed.

The relative performance of actual and nominal triads suggests an alternative to the specific group development process used in this experiment. Perhaps it would be best if developers first worked alone, developing their own spreadsheets separately in parallel. Afterward, members of a triad would compare their spreadsheets and see where there were differences. They could then discuss their

differences and produce a consensus spreadsheet. A pretest with MBA students working on a different and more complex task found that they were able to identify spreadsheet differences quickly and reliably.

However, an experiment with parallel development in programming (Knight and Leveson 1986) suggests that developers tend to make mistakes in the same places, and this would reduce the benefits of parallel development. In addition, based on Panko and Halverson's observational experiences, developers may not be able to resolve differences effectively. Another concern is that if designs are very complex, they may be so different that comparing them could prove very difficult.

In addition, group *development* is not the only possible type of spreadsheet teamwork. Programmers often use team code inspection *after development* to look for errors in a program module. Team code inspection in software development catches about 60% to 80% of all errors (Panko 2001a), which is about the improvement rate seen in this group development study. In a spreadsheet code inspection experiment that used groups of three students, Panko (1999) found an 83% error detection rate for a spreadsheet seeded with errors. Code inspection may be less expensive than group development and so should be considered an alternative to group development. However, Beck (2000) notes that code inspection is viewed as painful by developers and tends to be resisted, while pair programming is widely accepted and viewed as pleasant. This difference could have strong implications for acceptance.

VII. REFERENCES[1]

Beck, K. *Extreme Programming Explained: Embrace Change*, Reading, MA: Addison-Wesley, 2000.

Brown, I. D. "Drivers' Margins of Safety Considered as a Focus for Research on Error," *Ergonomics* (33:0/11), 1990, pp. 1307-1314.

Brown, P. S., and Gould, J. D. "An Experimental Study of People Creating Spreadsheets," *ACM Transactions on Office Information Systems* (5:3), 1987, pp. 258-272.

Butler, R. "Is this Spreadsheet a Tax Evader? How H. M. Customs & Excise Test Spreadsheet Applications" *Proceedings of the Thirty-Third Hawaii International Conference on System Sciences*, Maui, Hawaii, January 2000

Butler, R. Personal communication, August and September, 1996. Data collected, 1992.

Chan, Y. E., and Storey, V. C. "The Use of Spreadsheets in Organizations: Determinants and Consequences," *Information & Management* (31:3), December 1996, pp. 119-134.

Coopers & Lybrand. **http://www.planningobjects.com/jungle1.htm**. Contact information is available at that webpage.

Cragg, P. G., and King, M. "Spreadsheet Modelling Abuse: An Opportunity for OR?," *Journal of the Operational Research Society* (44:8), August 1993, pp. 743-752.

Davies, N., and Ikin, C. "Auditing Spreadsheets," *Australian Accountant*, December 1987, pp. 54-56.

Ferdinand, A. E. *Systems, Software, and Quality Engineering: Applying Defect Behavior Theory to Programming,* New York: Van Nostrand Reinhold, 1993.

Gable, G.; Yap, C. S.; and Eng., M. N. "Spreadsheet Investment, Criticality, and Control," *Proceedings of the Twenty-Fourth Hawaii International Conference on System Sciences,* Vol. III, Los Alamitos, CA: IEEE Computer Society Press, 1991, pp. 153-162.

Galletta, D. F.; Abraham, D.; El Louadi, M.; Lekse, W.; Pollailis, Y. A.; and Sampler, J. L. "An Empirical Study of Spreadsheet Error-Finding Performance,"

---

[1]Editor's Note: The following reference list contains hyperlinks to World Wide Web pages. Readers with the ability to access the Web directly or are reading the paper on the Web can gain direct access to these linked references. Readers are warned, however, that

1. these links existed as of the date of publication but are not guaranteed to be working thereafter.
2. the contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. the author(s) of the Web pages, not AIS, is (are) responsible for the accuracy of their content.
4. the author(s) of this article, not AIS, is (are) responsible for the accuracy of the URL and version information.

Journal of Accounting, Management, and Information Technology (3:2), April-June 1993, pp. 79-95.

Galletta, D.F.; Hartzel, K. S.; Johnson, S.; and Joseph, J. L. "Spreadsheet Presentation and Error Detection: An Experimental Study," *Journal of Management Information Systems* (13:3), Winter 1996-1997, pp. 45-63.

Hall, M. J. J. "A Risk and Control Oriented Study of the Practices of Spreadsheet Application Developers," *Proceedings of the Twenty-Ninth Hawaii International Conference on Systems Sciences, Vol. II,* Kihei, Hawaii, Los Alamitos, CA: IEEE Computer Society Press, January 1996, pp. 364-373.

Hassinen, K. *An Experimental Study of Spreadsheet Errors Made by Novice Spreadsheet Users,* Department of Computer Science, University of Joensuu, P. O. Box 111, SF-80101 Joensuu, Finland, 1988.

Hassinen, K. Personal communication, 1995.

Hicks, L. Personal communication, June 21, 1995.

Knight, J. G., and Leveson, N. G. "An Experimental Evaluation of the Assumption of Independence in Multiversion Programming," *IEEE Transactions on Software Engineering* (SE 12:1), 1986, pp. 96-109.

KPMG Management Consulting. "Supporting the Decision Maker: A Guide to the Value of Business Modeling," press release, July 30, 1998, **http://www.kpmg.co.uk/uk/services/manage/press/970605a.html**

Lukasik, T., CPS. Personal communication, August 10, 1998.

Marquart, D. I. "Group Problem Solving," *Journal of Social Psychology* (4:1), 1955, pp. 103-113.

Nardi, B. A., and Miller, J. R. "Twinkling Lights and Nested Loops: Distributed Problem Solving and Spreadsheet Development," *International Journal of Man-Machine Studies* (34:1), 1991, pp. 161-168.

Nosek, J. T. "The Case for Collaborative Programming," *Communications of the ACM* (41:3), 1998, pp. 105-108.

Panko, R. R. "A Human Error Website," 2001a (**http://panko.cba.hawaii.edu/humanerr/**).

Panko, R. R. "Spreadsheet Research (SSR) Website," 2001b (**http://panko.cba.hawaii.edu/ssr/**).

Panko, R. R. "Applying Code Inspection to Spreadsheet Testing," *Journal of Management Information Systems* (16:2), 1999, pp. 159-176.

Panko, R. R., and Halverson, Jr., R. H. " Are Two Heads Better than One? (At Reducing Errors in Spreadsheet Modeling)," *Office Systems Research Journal* (15:1), 1997, pp. 21-32.

Panko, R. R., and Sprague, Jr., R. H. "Hitting the Wall: Errors in Developing and Code-Inspecting a 'Simple' Spreadsheet Model," *Decision Support Systems* (22), 1998, pp. 337-353.

Putnam, L. L. "Conflict in Group Decision Making" in R. F. Hirokawa and M. S. Poole (Eds.) *Communication and Group Decision Making,* Beverley Hills, CA: Sage, 1986, pp. 175-196.

Reason, J. *Human Error,* Cambridge, England: Cambridge University Press, 1990.
Shanteau, J., and Phelps, R. H. "Judgment and Swine: Approaches and Issues in Applied Judgement Analysis," in M. F. Kaplan and S. Schwartz (eds.), *Human Judgment and Decision Processes in Applied Settings*, New York: Academic Press, 1977.
Steiner, I. D. *Process and Productivity,* New York: Academic Press, 1992.
Stroebe, W., Diehl, M., and Abakoumkin, G. "The Illusion of Group Effectivity," *Personal and Social Psychology Bulletin* (18), 1992, pp. 643-650.
Wagenaar, W. A., and Keren, G. B. "Does the Expert Know? The Reliability of Predictions and Confidence Ratings of Experts," in E. Hollnagel, G. Manici, and D. D. Woods (Eds.), *Intelligent Decision Support in Process Environments*, Berlin: Springer-Verlag, 1986, pp. 87-103.
Wheelan, S. A., and McKeage, R. L. "Development Patterns in Small and Large Groups," *Small Group Research* (24:1), 1993, pp. 60-83.
Williams, L. A. *The Collaborative Software Process*, Unpublished Ph.D. Dissertation, Department of Computer Science, University of Utah, 2000.

VIII. ABOUT THE AUTHORS

Raymond R. Panko is a professor of business administration at the University of Hawaii. He received his doctorate from Stanford University and was a project manager at SRI International before coming to the University. He has been conducting research on end user computing since the 1960s and on spreadsheet errors and other information technology risks since the early 1990s. His home page is **http://panko.cba.hawaii.edu/**.

Richard P. Halverson, Jr. is president of Guide.Net, Inc., in Honolulu, Hawaii. He received his Ph.D. from the University of Hawaii in Communication and Information Sciences.

Journal of the Association for Information Systems

JAIS EDITORIAL BOARD

ADMINISTRATIVE PERSONNEL