AD_____

MIPR NUMBER 95MM5576

TITLE: Instructional Interventions for Reduction of Gender Differences in Learning

PRINCIPAL INVESTIGATOR: J. Wesley Regian, Ph.D.

CONTRACTING ORGANIZATION: Armstrong Laboratory/CFT
Brooks AFB, Texas 78235-5241

REPORT DATE: April 1996

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1996	Final (9 Jan 95 - 30 Sep 95)

4. TITLE AND SUBTITLE
Instructional Interventions for Reduction of Gender Differences in Learning

5. FUNDING NUMBERS
95MM5576

6. AUTHOR(S)
J. Wesley Regian, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Armstrong Laboratory/CFT
Brooks AFB, Texas 78235-5241

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

14. SUBJECT TERMS Defense Women's Health Research Program

15. NUMBER OF PAGES
70

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

__✓__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

4 Apr 96

J. WESLEY REGIAN, Ph.D. Date

TABLE OF CONTENTS

ACKNOWLEDGMENTS

We would like to offer special thanks to: Wes Regian, who co-wrote the original proposal of this research; Barry Goettl, who greatly assisted in preparing the procedural learning task; Cathy C. Gomez, who provided valuable suggestions about aspects of this paper; and Pat Kyllonen, who assisted in some of the statistical analyses. We'd additionally like to thank Bill Uhland, who performed the testosterone radioimmunoassay procedure on the saliva samples, and Eddie Gomez, who served as the medical monitor for the experiments involving testosterone. Finally, we need to acknowledge the great assistance provided by Linda Robertson-Schüle, Rickard Robbins, Shirley Snooks, Jason Miller, and Kevin Gluck in the massive data collection job. Finally, we offer our appreciation to Kevin Kline and Pamela Goettl for their important assistance across all of these experiments.

The research reported in this paper was conducted as part of the Armstrong Laboratory, Human Resources Directorate. This study represents basic research funded by the Defense Women's Health Research Program. The opinions expressed in this article are those of the authors and do not necessarily reflect those of the Air Force. Correspondence concerning this paper should be addressed to Valerie J. Shute, Armstrong Laboratory (AL/HRTI), 1880 Carswell Avenue, Lackland Air Force Base, Texas, 78236-5507.

The Effects of Instructional Intervention, Gender, Testosterone, and Stress on Spatial Learning

Many studies in the literature have reported gender differences in spatial abilities (e.g., Linn & Peterson, 1985; Maccoby & Jacklin, 1974; McGee, 1979; Voyer, Voyer & Bryden, 1995). Invariably, these studies have shown that males perform better on a wide array of small-scale spatial tasks compared to females. Furthermore, there appears to be a direct relationship between spatial skills and endogenous testosterone level (Kimura, 1992; Shute, Pellegrino, Hubert & Reynolds, 1983). The nature of this relationship is such that females with higher levels of testosterone perform better on spatial tasks than females with lower levels of testosterone, while males with low to moderate levels of testosterone perform better than males with excessively high levels of testosterone. More complex spatial tasks involving dynamic components have also been reported that show the same male superiority with regard to differential performance (e.g., Shebilske, Regian, Arthur & Jordan, 1992).

The objectives of the studies reported herein sought to refine and extend earlier research on gender-based spatial differences, focusing on the application of intructional interventions that can possibly reduce any gender differences. First, we attempted to replicate a study where one simple instructional intervention (i.e., heterogeneous discussion groups between training sessions) had been shown to improve females' spatial skills. Second, we tried to account for the remaining performance differences between males and females after the instructional intervention by measuring testosterone levels in saliva using a radioimmunoassay procedure. Third, we measured gender-based differences in performance on a complex spatial task under stress, and related these differences to testosterone level.

Experiment 1

In the following experiment, we sought to replicate findings reported by Regian & Shute (1993). Specifically, that study involved the use of an instructional intervention that took the form of brief discussion groups (5 min. in duration and following specific training sessions). Results showed that female performance was dramatically enhanced by this instructional intervention, but there was no effect on male performance (perhaps due to ceiling effects). Consequently, in this study, we hypothesized that there would be a main effect due to discussion group, particularly for the females. In addition, we wanted to account for remaining gender differences on the task via endogenous testosterone levels, and examine the nature of the relationship between testosterone level and spatial skill. Spatial performance in this experiment was defined as the final outcome score for Space Fortress (Session 10), holding baseline score constant.

Method

Participants. Two hundred persons (131 males and 69 females) participated and completed Experiment 1. All participants were recruited through local temporary agencies in San Antonio, Texas and paid approximately $5.00 per hour for their involvement. Participants ranged in age from 18 to 30 years (M = 22.1, SD = 3.5), and had a high school diploma or GED, but had not completed a four-year college degree. Screening was conducted to eliminate those persons with prior exposure to the task (Space Fortress) and those who could not obtain a score of at least 780 on the Space Fortress Aiming Task (Mane & Donchin, 1989) administered on the first day. In addition, participants who reported playing more than 20 hours of video games per week were excluded from the study.

Equipment. The experiment was conducted in the USAF Armstrong Laboratory located at Lackland Air Force Base, Texas. The laboratory consists of 30 Compaq 486/33L computers with NEC/Multisync SVGA monitors. Learners used standard keyboards, CH Products FlightStick joysticks and Logitech mice.

Task. The version of the Space Fortress task that was employed in this study was similar to the one used by Gopher, Weil, and Bareket (1992), which represents a modified version of the program used by Mane and Donchin (1989). In Space Fortress, learners use a joystick to fly their ship around a fortress and try to destroy it while avoiding being shot by the fortress or destroyed by mines. Several secondary tasks exist, such as being able to identify something as friend or foe, and selecting bonus opportunities.

The overall score learners receive is based on the summation of four subscores involving speed, velocity, points, and control components. In the modified version of the program that we used, all four subscores are displayed on the screen (see Gopher, et al., 1992 for a complete description of the task). The task is extremely complex and learners typically score as low as -3000 points during the initial training (baseline) session.

Training sessions consist of ten 3-minute standard game trials. During each of the subsequent trials, if the learner's ship is destroyed, it is repositioned at the starting point and the trial continues. Learners receive performance feedback (i.e., subscores and total score) at the end of each three-minute trial. Learners are instructed that during each module, the first eight games are practice games and the final two games are test games.

Procedure. The design was a 2×2 factorial: *instructional intervention* (individual vs. group discussion) $\times$ *gender* (male vs. female). Participants were run in groups of 25 across a 2-day period. On the first day of each administration, they reported to the laboratory, were given an

overview of the experiment and general procedures, signed a consent form, and were given the Space Fortress screening task. Participants received only three 15-minute breaks (one in the morning and two in the afternoon) and one 30-minute lunch period each day. Saliva was collected at approximately 10:00 a.m. on day 1 of the experiment for each group. A radioimmunoassay (RIA) procedure was performed on the saliva samples to measure participants' endogenous testosterone levels, in duplicate (see Testosterone Collection Protocols below for more information).

Instructional Intervention. The discussion group protocol consisted of randomly assigning learners to either an individual or discussion group condition (the latter consisting of two males and two females). Following the video presentation of task instructions and the four-game baseline, each discussion group was asked to arrange their chairs together in a semi-circle alternating in a male-female fashion. Discussion groups were positioned in remote areas of the room by the lab administrators so that individual learners would not be able to overhear any of the groups' discussions concerning the task. Prior to the start of each discussion session, learners were given the following instructions: "You will now have 15 minutes to discuss how to perform the Space Fortress task as efficiently as possible. Discuss what you can do to improve your performance, or things you have learned that can increase your score. The four of you will meet several times to discuss Space Fortress." A total of five discussion sessions were held following the baseline, 1st, 3rd, 5th, and 7th training sessions.

The individual learners remained at their desks and were given instructions for an anagram task. They were asked to do their best at creating as many words from the letters of the given words, and told it would be used as an index of their verbal ability. All participants (individual and discussion group learners) were given strict instructions not to discuss the

training task, their strategies, or their performance scores at any time other than when they were placed into their timed discussion group sessions. If they had not been placed in a discussion group, they were not allowed to talk about the task, their strategies, or scores to anyone.

Testosterone Collection Protocols. Upon recruitment by the temporary agencies, participants were told that a saliva sample would be collected and tested for endogenous levels of testosterone. When they arrived at the laboratory for training, they were given a voluntary informed consent statement to sign (see Appendix 1) indicating the procedures, the reason for the sample collection, and any possible risk involved in the procedure. During the initial briefing of general instructions, participants were asked that they refrain from smoking, eating, drinking (except water), or chewing gum for the next hour. This was done to prevent a contamination of the saliva sample. Approximately 10-15 ml of saliva were collected in test tubes from each person. These samples were appropriately labeled and refrigerated for 24 hours. Samples of participants that did not return for the second day of testing were discarded. The RIA method used to determine the testosterone levels is described in Appendix 2.

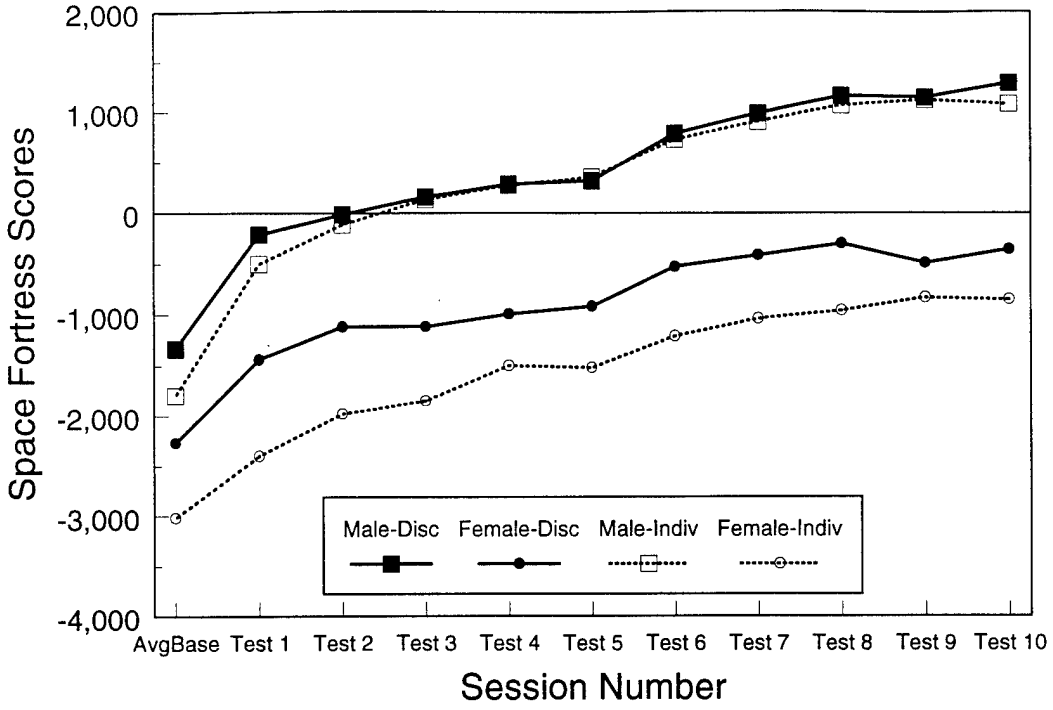
Results

As with almost all previous studies that have employed the Space Fortress task, this study similarly showed incoming gender differences among learners, as measured by a four-game baseline session. An ANOVA was computed on the baseline data (the average of 4 trials) by gender. The results showed that there was a significant main effect due to gender: $F(1, 196) = 27.88$, $p < .001$. Males started the task with higher average baseline scores ($M = -1631.31$, $SD = 1257.39$) compared to females ($M = -2509.34$, $SD = 1390.38$). We also wanted to test incoming gender differences related to demographic variables (i.e., age and video-game experience). We computed a MANOVA on these two variables by gender, and there were no significant

differences found: Wilks' exact $F(2, 176) = 0.96$, $p = .43$. Thus the obtained gender differences in incoming skill were not attributable to either of these variables.

To answer the question of whether males and females differ in their relative acquisition of this complex spatial task, and also whether females benefit more by participating in group discussions compared to males, we computed a repeated-measures ANOVA with *session* as the within-subject variable (i.e., performance data from Sessions 1 to 10) and *gender* and *instructional intervention* (group vs. no group discussion) as the between-subject variables. We included participants' average baseline score as a covariate in the equation.

Results showed a main effect due to *session*: $F(9, 187) = 21.81$, $p < .001$. Overall, learners improved across the 10 learning sessions. Next, there was a significant main effect of *gender*: $F(1, 194) = 19.82$, $p < .001$, whereby males consistently outperformed females across all sessions. There was no main effect due to *instructional intervention:* $F(1, 194) = 0.03$, $p = .87$, nor an *intervention* × *gender* interaction: $F(1, 194) = 0.94$, $p = .33$. The learning data across the ten sessions, divided by instructional intervention and gender, are shown in Figure 1.



Notes. Discussion groups met after the following training sessions: Baseline, 1, 3, 5, and 7. Sample sizes per condition were: Male-Disc (N = 48); Female-Disc (N = 47); Male-Indiv (N = 82); Female-Indiv (N = 22)

Figure 1. Performance across 10 training sessions by gender and instructional intervention.

Testosterone Level. Our next research question examined individuals' testosterone levels to see if the hypothesized quadratic relationship between testosterone level and spatial skill was upheld. That is, for females, more testosterone was hypothesized to be associated with greater spatial performance, and for males, low to moderate levels were hypothesized to be optimal. Participants' testosterone levels were assessed via a radioimmunoassay procedure using saliva samples that were divided into two separate samples. Duplicate assays were run on both saliva samples, and there was a significant correlation between the two assays ($r_{xy} = 0.86$ for samples 1 and 2, N = 199). The average of the two samples comprised our testosterone value used in all subsequent analyses.

First, we found significant differences between gender with regard to testosterone level: Male $\underline{M} = 71.2$ pg/ml, Female $\underline{M} = 20.1$ pg/ml, $\underline{F}(1, 197) = 127.82$, $\underline{p} < .001$. Thus, the RIA

method reliably distinguished between gender. Second, to test the hypothesized quadratic relationship between spatial skill and testosterone level, we computed a regression analysis--curve estimations for linear, quadratic, and cubic trends--predicting outcome performance (i.e., session 10 test score). All three trends fit the data well, and the results from the quadratic fit alone was: $R^2 = 0.19$, $\underline{F} = 15.06$, $\underline{p} < .001$. A scatterplot of all three trends is shown in Figure 2.

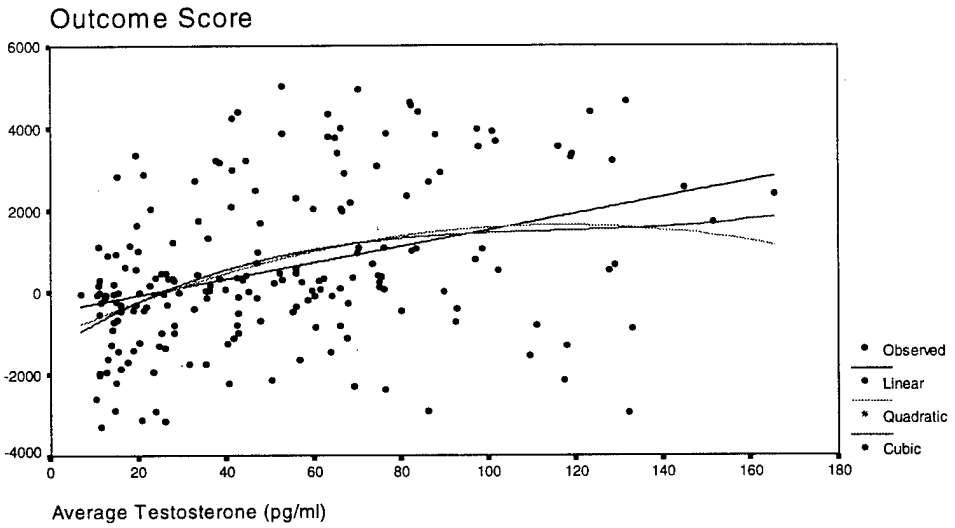


Figure 2. Linear, quadratic, and cubic fits of testosterone level by spatial outcome performance.

Effects of gender, testosterone level, and instruction on Space Fortress performance.

Consistent with past research, we found that gender differences exist throughout training on the Space Fortress task. However, we failed to find a main effect due to instructional intervention. A related series of questions concern interactions involving gender, instructional intervention, and testosterone level. For example, testosterone level may be more predictive of outcome performance for males but not females. Alternatively, instructional intervention may be useful for females but not males. To address these questions, we computed a regression analysis predicting final performance score on session 10. The independent variables were gender, instructional treatment, testosterone level, and the interactions among these variables.

First, we examined the interactions between: (a) instructional treatment and gender, (b) instructional treatment and testosterone, and (c) instructional treatment and the gender by testosterone interaction (i.e., a three-way interaction). None of these interactions were significant (all three F's < 1). This suggests that the instructional intervention did not in any way mediate the relationship of gender and testosterone with Space Fortress performance.

Next, we examined the relationship between gender and testosterone with regard to predicting Space Fortress performance. A significant interaction here would indicate that testosterone levels might affect males and females differently; for example, testosterone might affect males' performance on Space Fortress but not females'. This analysis examined the contribution of the gender by testosterone product variable, with (a) instructional intervention, (b) gender main effect, and (c) testosterone level already in the model. The results of this analysis showed that there was no interaction between testosterone level and gender, $\underline{F} = 1.02$, $\underline{p} > .10$.

Because the previous two regression analyses suggested that gender, testosterone, and instructional treatment did not interact, the final model simply included the main effects of gender, testosterone level, and treatment. In this analysis there was a significant effect for both gender, $\underline{t} = 3.63$, $\underline{p} < .01$, and testosterone levels, $\underline{t} = 2.35$, $\underline{p} < .05$, suggesting that testosterone level does affect Space Fortress performance, even controlling for gender. It also suggests that at any given testosterone level, males will perform better on Space Fortress than females. In this analysis, instructional intervention did not have any effect on performance. A plot of the final model is shown in Figure 3.

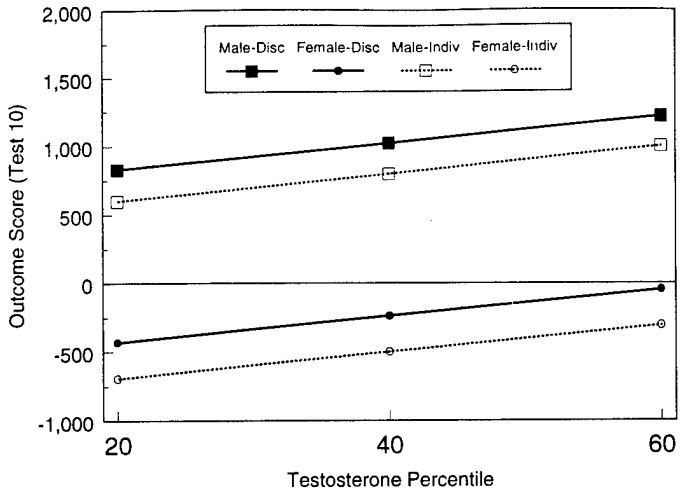


Figure 3. Main effects of gender, testosterone level, and treatment on final outcome score.

Discussion

This experiment attempted to replicate a study where brief discussion groups between training sessions dramatically enhanced female performance (Regian & Shute, 1993), and then account for any remaining gender differences with endogenous testosterone levels. Learners were placed in intermittent discussion groups (two males and two females) between training sessions to share experiences and tactics for improving their scores on the spatial task. We did find (a) main effects of testosterone level on task performance (i.e., a significant quadratic relationship between testosterone level and spatial skill), (b) main effect of gender (i.e., males consistently performed better on the task than females, across all sessions), (c) no main effects of instructional intervention, and (d) no interactions among the independent variables.

We were surprised to find no significant effect of instructional intervention on task performance (particularly for the females). However, females placed in discussion groups performed better than females working independently, but their skill level never reached that of independent males. Also, females in the discussion group began the task with slightly higher incoming scores (by chance) than their counterparts in the individual condition. Males in both

conditions (discussion and individual) performed equivalently, and significantly better than females throughout training. Overall, females never scored above 0, suggesting that there may be individual component skills of this task that were simply not being acquired (e.g., controlling velocity, circumnavigating the frictionless space). Our future research plans involve investigating these component skills to determine if any relationships exist. We now turn our attention to another variable that may impact complex spatial performance--stress. In particular, we test for differences in learning and performance as a function of stress and its interaction with gender.

Experiment 2

Feelings of stress have been shown to differentially impact learning and/or performance, and not always in a negative sense. Just as people may differentially respond to various types of instructions, a considerable body of extant literature suggests that there are many reliable individual differences in responses to various types of induced stress (Eysenck, 1983; Jones, 1983; Spielberger, Gorsuch, & Lushene, 1970).

Depending on the nature and degree of the stress manipulation, a quadratic trend has been reported in the literature whereby moderate levels of stress may be optimal in relation to performance. As a classic example, Yerkes and Dodson (1908) found a relationship between arousal/stress and performance. Foot shocks were administered while learning a visual discrimination task, which ranged from easy to difficult. When the task was easy, increasing the shock level (and thus the stress level) actually increased performance on the task. But when the task became more difficult, a negative relation was found between shock level and performance. Optimal performance was associated with moderate levels of foot shocks.

In general, the literature in this area is unclear about the possible interaction between stress and gender. Some researchers (Spielberger et al., 1970) suggest that females are more

emotionally labile than males in their reactions to highly stressful or relaxing circumstances. Spielberger and colleagues made this conclusion based on a study where learners were first shown a stressful movie (wood-shop accidents), then either given stress training or not. Females reported more perceived stress compared to males based on the State Anxiety Inventory. However, the females that had been given stress-reduction training also indicated less stress on the inventory than males that had been given the same training.

Although the literature suggests that the relationship between stress and learning is different for men and women, the literature is not clear on what types of manipulations are best for examining stress. Two main stress variables were examined in this experiment--sound and task instability. The purpose of this experiment was to select a robust and reliable stress variable to employ in subsequent, larger studies.

In the *sound* condition, a continuous background noise (cafeteria chatter) was played on two strategically-placed portable stereos, at a level of 80 dB during the performance of the task. The intention was that this auditory input would disrupt attention to the task, thus serving as a source of stress. The control condition was simply one of no sound (i.e., the normal task). All participants in the sound condition were tested as a group, at the same time.

The *task-instability* condition involved a manipulation where individuals were told that certain aspects of the task would be changed (for the upcoming sessions 11, 12, and 13), but were not told the nature of the alterations, or how to deal with them. In reality, however, no changes to the task were implemented (hence, invoking uncertainty which was believed to induce stress).[1]

[1] Originally, we planned to actually change the task in accord with the instructions given to the participants (cited above). However, during this experiment, the intended changes were not implemented by the program, thus no changes actually occurred. We didn't mean to lie to the subjects, but this "fake out" resulted in some very intriguing results which made this accidental condition an excellent stress manipulation for Experiment 3.

The control condition was simply a normal, stable version of the task, where participants were not informed of any changes, nor were any changes in the task presented.

Method

Participants. Fifty-seven individuals (37 males and 20 females) participated and completed Experiment 2. Similar to Experiment 1, all participants were recruited through local temporary agencies in San Antonio, Texas, paid approximately $5.00 per hour for their participation, ranged in age from 18 to 30 years (M = 22.1, SD = 3.4), and had a high school diploma or GED (but not a four-year college degree). Again, screening eliminated those with prior exposure to the task (Space Fortress) and those who could not obtain a score of at least 780 on the Space Fortress Aiming task (Mane & Donchin, 1989). In addition, persons who reported playing more than 20 hours of video games per week were excluded from the study.

Procedure. The design was a 2 × 2 × 2 factorial: *sound* (no sound vs. sound) × *instability* (stable vs. unstable task) × *gender* (male vs. female). The experiment was conducted subsequent to Experiment 1 to examine various stress manipulations such that an appropriate stress variable could be identified for use in subsequent studies. Learners were trained on the Space Fortress program in groups of 25, across a 2-day period. Following 10 training sessions (on day 1), learners were then randomly assigned to one of the four treatment conditions: (a) sound-stable, (b) sound-unstable, (c) no sound-stable, or (d) no sound-unstable. Participants then proceeded to complete three additional training sessions (11, 12, 13).

Heart rate was measured before and after each of the three Space Fortress sessions for a physiological measurement of stress. Learners were instructed on the correct procedure to measure their own heart rate (i.e., locate their pulse on their wrist or neck and count the number of heartbeats within a specified time). Lab administrators timed each pulse-reading session for a

duration of 10 seconds. Individuals recorded their own values and later multiplied it by 6 to arrive at a measure of heartbeats per minute. These final values were used in subsequent analyses. The State-Trait Anxiety Inventory for Adults (form Y-1, Speilberger, 1983) was used to measure individuals perceived stress state. This inventory was administered on the computer immediately following each of the new training sessions (i.e., 11 - 13). Once learners completed the inventory, they proceeded to the next Space Fortress session.

Our primary question concerned the influence of different kinds of stress manipulations (i.e., background sound and task instability) on three different dependent measures: (a) *Performance*--Space Fortress performance across three sessions, (b) *Heart rate*--pulse-rate measurements assessed by heartbeats per minute, and (c) *Self-reports* of perceived stress from a State-Trait Anxiety questionnaire for adults.

Results

Data Reduction. To reduce and simplify the data from our dependent variables, across the three sessions, we computed a factor analysis (principal axis factoring, varimax rotation) on the performance, heart rate, and self-report data, collectively. The results of this analysis were the extraction of three orthogonal factors: Performance, heart rate, and self-report. Factor scores were saved for each person on each of these three factors. The percentage of variance accounted for by the three factors was 88.5%, and each factor had $\underline{M} = 0$, and $\underline{SD} = 1$. Finally, factor loadings, per factor, were all high: (a) Performance factor loadings ranged from .93 - .96, (b) heart rate loadings ranged from .86 - .91, and (c) self-report loadings ranged from .93 - .95. These factor scores were used as the dependent measures in subsequent analyses.

A MANOVA was computed on the dependent measures (performance, heart rate, and self-report factor scores), with the following between-subject variables: *sound* (coded as 0, 1 for

no sound and sound, respectively), *instability* (coded as 0, 1, for normal and unstable task, respectively), and *gender* (coded as 0, 1 for males and females, respectively). Results of the analysis showed the following. There was no significant main effect on the dependent measures due to *sound*: Wilks' exact $\underline{F}(3, 37) = 2.33$, $\underline{p} = .09$, nor was there a main effect attributable to *gender*: Wilks' exact $\underline{F}(3, 37) = 1.46$, $\underline{p} = .24$. There was, however, a significant main effect due to *instability*: Wilks' exact $\underline{F}(3, 37) = 3.85$, $\underline{p} = .02$. Univariate F-tests showed that this was solely a function of its effect on the performance factor: $\underline{F}(1, 39) = 9.95$, $\underline{p} = .003$. With regard to the standardized performance scores, the data were surprising in that they showed much higher scores associated with the unstable condition ($\underline{M} = .65$) compared to the normal condition ($\underline{M} = -.30$). This suggests that the element of task unpredictability enhanced performance relative to the control condition, perhaps serving to arouse participants, thereby making them more conscious of the task and thus enhancing performance.

While none of the two- or three-way interactions were significant, the interaction between sound × instability approached statistical significance: Wilks' exact $\underline{F}(3, 37) = 2.59$, $\underline{p} = .07$, and the univariate F-test for the performance factor showed this interaction to be significant: $\underline{F}(1, 39) = 5.14$, $\underline{p} = .03$. The interaction is shown in Figure 4. Briefly, the highest performance scores were associated with the most "stressful" condition: sound and unstable task.

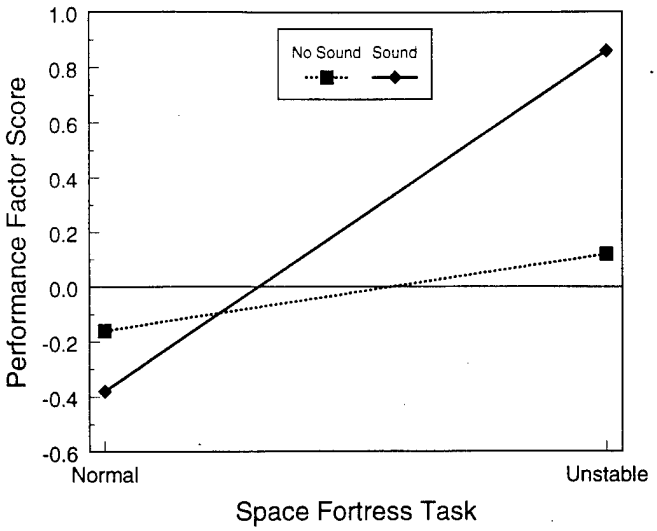


Figure 4. Interaction between task and sound conditions on performance factor score

Discussion.

Our goal in conducting this relatively small experiment was to ascertain a good stress variable that would allow us to examine gender differences in relation to stress during the acquisition of a complex spatial task. Our dependent measures did not differ significantly due to the sound variable, although results approached statistical significance ($p = .09$). Overall, participants in the sound condition actually showed higher scores ($M = 0.43$) compared to the no-sound condition ($M = -0.08$). Thus, learners' performances appeared to be enhanced when background sound was played compared to the no-sound condition. It is likely that the sound manipulation may have actually served to arouse learners (i.e., a positive mediator), rather than stress them (i.e., a negative mediator).

This assumption, initially posited by Yerkes & Dodson (1908), is further reinforced by the fact that the instability variable evidenced the same pattern, and showed significant differences in relation to the dependent measures. That is, rather than inhibiting performance, participants in the unstable-task condition performed superior to those in the normal task

condition. Furthermore, the heart-rate data (although not statistically significant) showed that learners in the unstable condition also had faster (standardized) heart rates ($\underline{M}$ = 0.27) compared to those in the normal task condition ($\underline{M}$ = 0.06). Finally, the interaction depicted in Figure 4 further supports the notion that our selected "stress" conditions were actually arousing learners. Thus, these manipulations may have related to performance in a positive linear function (i.e., falling on the upward slope of the aforementioned quadratric trend). Had we employed a range of sounds (e.g., > 80 decibels) or introduced greater task instability (e.g., stating that certain changes would occur then having the opposite occur), then we may have obtained results that served to decrement performance. This hypothesis will motivate future research.

Taken together, these data suggest that background sound and task instability, as defined within our paradigm, actually had a tendency to arouse, rather than stress learners. Gender differences were not seen on the dependent measures suggesting that males and females dealt with the "stress" variables similarly. For example, heart rates and self-reports of stress were comparable between males and females across the different conditions. Given the significant findings of the task-instability variable (and the non-significant findings related to sound), this is ultimately the stress manipulation we selected for additional testing in Experiment 3.

Experiment 3

Results from Experiment 2 suggested that the manipulation of an uncertainty variable yielded differences in relation to the dependent measures. Further, we showed that the variable actually appeared to benefit performance, overall. This finding is consistent with previous research which suggests that small to moderate amounts of stress may actually elevate performance by arousing learners (e.g., Yerkes & Dodson, 1908). The primary goal of Experiment 3 was to further examine the effects of stress/uncertainty on performance, and further evaluate the stress by gender/testosterone interaction in terms of differential performance on a complex spatial task.

As mentioned earlier, the stress manipulation in Experiment 2 (instability of task) actually represented a mistake in the program. That is, changes to the task components that were supposed to occur never actually took place. This discrepancy between what learners were told and what actually transpired was believed to cause some degree of cognitive uncertainty, and hence stress. For the current experiment, we corrected this program flaw, but retained the unstable condition. This resulted in three treatment conditions. If performance is enhanced, rather than inhibited, by elevating uncertainty/stress, then the order of performance scores by condition is hypothesized to be: unstable > normal > altered. The rationale for this ordering is as follows. First, learners have acquired, via ten preceding sessions, the prerequisite skills at the outset of this experiment (i.e., session 11). When learners in the unstable condition are informed that the task will subsequently change in some ways, this is expected to invoke cognitive arousal. However, because no changes actually occur, their performance (given the elevated arousal) is expected to be higher compared to the "normal" condition. On the other hand, the "altered" condition really does implement changes to the task so learners have to acquire several new rules

and procedures for successful performance. Thus, their performance is expected to be less than the "normal" condition given the new learning and skills that must be acquired. Finally, with regard to any interactions between instability condition and gender, we hypothesized the aforementioned ordering of conditions, but only for the females who tend to be more susceptible to changes in task characteristics compared to males, who are, overall, more facile in performing the task.

To assess the role of testosterone with regard to performance on the spatial task under conditions of stress, we obtained two saliva samples from each participant prior to the experiment, and conducted a radioimmunoassay of testosterone level from their samples. Because of the exploratory nature of this study, we divided male and female data (separate analyses) into high/low categories based on median splits of their testosterone level. This enabled us to test hypotheses related to differential reactions to stressful manipulations as a function of relative testosterone level. We speculated that high-testosterone females would be less affected by unstable task characteristics compared to low-testosterone females, who in turn were believed to be maximally effected by task instability, in a negative sense. Thus, high-testosterone females' relative performance on the spatial task under stress conditions was expected to exceed that of low-testosterone females, and be similar to the low-testosterone males. We did not explicitly posit any performance differences on the task between low- and high-testosterone males as both the linear and quadratic trends (from Experiment 1) were significant (i.e., more testosterone, overall, was associated with better performance on the spatial task, although a significant quadratic relationship between testosterone level and spatial performance was obtained). We were interested in testing for differences between the low- and high-testosterone males with regard to performance under stressful conditions.

Method

Participants. A total of 185 participants (117 males and 68 females) completed Experiment 3. These individuals also participated in the acquisition phase for Experiment 1 (10 training sessions on day 1), but were then exposed to various stress conditions during the subsequent phase (3 training sessions on day 2) for this investigation on stress, gender, and performance. See the *Participants* section in Experiment 1 for a description of these individuals.

Procedure. Experiment 2 consisted of only two treatment conditions related to the instability variable: normal and unstable task. For this experiment, we added a third condition resulting in two controls for the unstable-task condition, both representing a stable contrast to the uncertain/inconsistent one. The three conditions in Experiment 3 included: (a) *normal* task, where participants were not informed of any changes, nor were any changes in the task presented, (b) *unstable* task, where participants were told that changes would occur, but not told what the changes were, or how to deal with them (note: no changes to the task actually transpired), and (c) *altered* task, where, similar to the unstable condition, participants were informed that changes would occur and, in fact, the task did change. As with the unstable task, participants were not told of the nature of the changes, nor how to deal with them. Some examples of task components that were altered from what was previously learned included: changing the scoring associated with bonus points and increasing the frequency of mines appearing on the screen.

The design was a 3×2 factorial: *stress* condition (normal vs. unstable vs. altered) $\times$ *gender/testosterone* category (female-low, female-high, male-low, male-high). Following the acquisition phase (baseline through session 10 from Experiment 1), learners were assigned to one of three treatment conditions. Learners in all conditions were instructed not to discuss any

features of the task, or what they believed had changed, but were informed of these changes (or lack thereof) at the end of the experiment.

Results

To test the relationship between testosterone level and instability condition on task performance, we computed two median splits on testosterone levels, separately for the male and female data. This resulted in four testosterone categories: male (low and high) and female (low and high). Our dependent measure was Space Fortress performance score on the very first session involving the stress manipulation (i.e., session 11) because if any effects were to be found, they would show up in this initial session. Subsequent sessions (12 and 13) would show attenuated effects given additional practice and learning opportunities that ensued. Session 10 performance scores were included as a covariate in the equation to control for gender differences that we knew existed at the conclusion of Experiment 1: $F(1, 197) = 40.90$, $p < .001$, Male $M = 1163.61$ ($N = 130$), Female $M = -515.42$ ($N = 69$).

We computed an ANOVA on Session 11 test data, with *testosterone* category and *instability* condition as the two between-subject variables. Session 10 test data served as the covariate. Results of this analysis showed a main effect due to *testosterone* category: $F(3, 184) = 3.13$, $p = .03$, and the data were ordered as follows: *Female-low* $M = -717.7$ ($N = 34$), *Female-high* $M = -584.6$ ($N = 35$), *Male-low* $M = 441.9$ ($N = 65$) and *Male-high* $M = 1462.6$ ($N = 63$). This represents a positive linear function where more testosterone is associated with better performance scores. Second, there was a significant main effect due to *instability* condition: $F(2, 184) = 13.55$, $p < .001$. The order of this variables was, from lowest to highest: *altered* $M = -266.2$ ($N = 45$), *normal* $M = 506.2$ ($N = 72$), and *unstable* $M = 644.2$ ($N = 80$). Again, this was in line with what we described earlier regarding superior performance in the unstable condition

given the arousal hypothesis. Finally, the *testosterone × instability* interaction was significant: $F(6, 184) = 2.70$, $p < .02$. This relationship, shown in Figure 5, suggests that the performance of high-testosterone males is completely unaffected by the task condition under which they were assigned--they perform consistently high in all three conditions. However, low-testosterone males show a different pattern--better performance in the unstable condition relative to the normal one, but a dramatic decline in performance in the altered condition. High-testosterone female data mirror that of the low-testosterone male pattern with an elevation in scores for the unstable condition (relative to normal), and a large decline in performance for the altered condition. Finally, low-testosterone females perform optimally under the normal condition, and when instability is introduced (even though the task does not actually change), performance drops, and continues to decline when the task is actually altered.

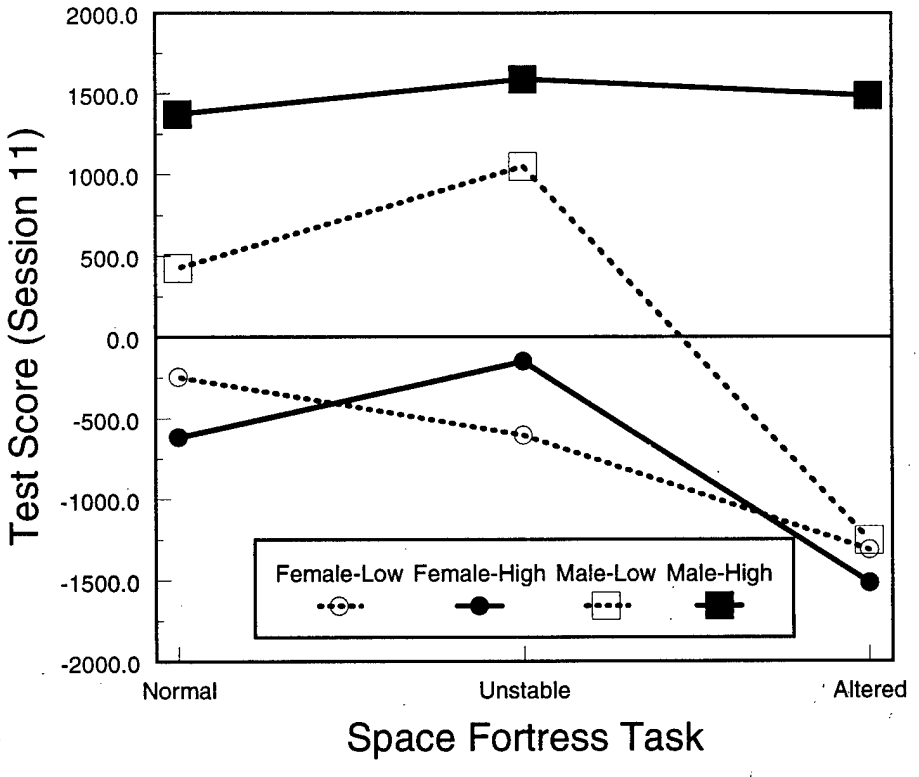


Figure 5. Interaction between task condition and testosterone level on performance score

Discussion

The goal of Experiment 3 was to further examine the relationship between testosterone (by gender) category and stress (task instability) on spatial performance. We succeeded in replicating the unexpected findings obtained in Experiment 2 concerning the superiority of the "unstable" condition relative to the "normal" one. Further, we postulated and confirmed that the "altered" task produced the poorest outcomes, overall. When we analyzed performance differences as a function of testosterone category and task characteristics, we again found main effects of testosterone level (more is better, overall), but of most interest was the obtained testosterone × task interaction.

We had predicted that low-testosterone males and high-testosterone females would perform comparably across the three different task conditions, and this hypothesis was confirmed. On the other hand, the high-testosterone male data were somewhat surprising in that they showed equivalent and high performance across all three task conditions. This finding can be explained in terms of the main effect of testosterone. That is, this group of learners acquired the complex spatial task quite readily, and when the task was rendered even more difficult (i.e., in the "altered" condition), this did not harm their performance. They simply acquired and integrated the new rules and continued to perform at a high level, in stark contrast to the other three groups of learners. The others (low-testosterone males and high- and low-testosterone females) were negatively affected by the altered task, most likely because they had not acquired the necessary skills associated with the original task sufficiently in the preceding 10 sessions.

Another interesting finding from this study showed how the low-testosterone males and the high-testosterone females performed similarly on the unstable task. They both showed higher

performance relative to the normal condition, suggesting that this condition served to arouse (and thus enhance) their performance. In contrast, the worse performance shown by the low-testosterone females in the unstable condition (relative to normal) suggests that the information they received about the "change of rules" (although the task did not actually change) was sufficiently stressful to significantly reduce performance.

In general, the obtained interaction may be explained in terms of relative task expertise, which in turn is mediated by testosterone level. Novices and experts differ in their perceptions of what constitutes a "stressful situation" based on their understanding of the task and the task demands. For the "unstable" condition, it is likely that males, along with high-testosterone females, readily determined that the task had not actually changed at all, despite reading the introductory statement to the contrary. On the other hand, low-testosterone females, still struggling with the complex spatial task, were not able to detect that the task had not changed. The implication of this finding is that some minimal level of proficiency must be obtained before introducing new task characteristics, even "fake" ones like were employed in the unstable task condition.

Summary and Conclusions

In general, the goals of these three experiments were to examine issues related to gender differences in complex spatial performance (under both normal and stressful conditions), seek ways to attenuate those differences via instructional interventions, and further account for differences in terms of testosterone level.

In Experiment 1, the objectives were twofold: (a) replicate a previously-obtained effect of a particular instructional intervention (i.e., small, mixed-gender discussion groups) that had been shown to dramatically enhance female spatial-skill acquisition, and (b) replicate the relationship

between testosterone level and spatial skill. With regard to the first goal, we failed to obtain either a main effect of instructional intervention, or the hypothesized intervention × gender interaction. We believe that the reason for this failure was due to methodological differences between our current study and the one we sought to replicate. That is, in the original study, discussion groups were both video-taped and more specifically structured. In our study (Experiment 1), we only video-taped the final (N = 25) group of participants. We expect that the video-taping that occurred in the original study made participants more focused and attentive ("putting their best face forward") during the discussion period. Without video-taping, it was easy for participants to become careless and off-task, talking about things that may not have been related to the task. This, in fact, was observed on several occasions. Second, the instructions given to the individuals in Experiment 1's discussion groups were not as explicit as they could have been (or in relation to the original study). For instance, we asked participants to simply, "Discuss what you can do to improve your performance, or things you have learned that can increase your score." They had fifteen minutes to talk, and they were mostly unsupervised during this time. In contrast, the discussion groups that yielded the original intervention × gender interaction (reported in Regian & Shute, 1993) consisted of specifically structured discussions (e.g., "What strategies did you use, how did they work, what were the results", and so on). Thus, the original study, employing both video-taping and more structured discussions, would have provided a considerably more fruitful environment for learners to share and pick up new knowledge, skills, and strategies, especially compared to the discussion sessions in the current study.

The second goal of Experiment 1 was to replicate a quadratic relationship between testosterone level and spatial skill. While there was a significant quadratic relationship obtained

(Multiple $\underline{R} = 0.44$), the linear function actually explained our data better. The most likely reason for this finding was that in our sample, we simply did not have males with excessively high testosterone levels. In previous studies that have analyzed testosterone levels in relation to spatial skill, there appears to be more variability in the male data. Further, it is the very high testosterone males that show poorer spatial performance (i.e., the downward slope of the quadratic trend) relative to the low and moderate level males.

The primary goal of Experiment 2 was to identify a valid stress manipulation so that in subsequent studies, we could test for gender differences on a complex spatial task, particularly in relation to stress. We tested two main stress conditions, sound and task instability, and found that there was a significant main effect on performance due to task instability. The most interesting finding from this study, however, was that our "stress" condition actually enhanced performance, overall, relative to a control condition. Moreover, the best performance was evidenced by learners in the double-stress condition--sound and instability. Learners in this condition performed at about 1.0 standard units above average (i.e., 1-sigma effect size). The most plausible explanation for this phenomenon is that we employed stress variables that served to capture our participants attention, but did not harm performance. We were constrained at the outset to limit our auditory stimuli to a maximum of 80 decibels for the sound condition. However, this level was not that distracting, and learners may easily have habituated to it over time. Had we used a higher volume (e.g., 120 dB), we believe that we then would have seen performance decrements. Similarly, our "instability" condition was relatively innocuous in that we informed participants that some changes to the task would occur, but no changes actually did take place. Most learners quickly determined that the task was the same as it had been, so it did

not disrupt performance at all. In fact, performance was enhanced in this condition, presumably because learners were more cognitively alert.

Experiment 3 examined stress, testosterone category, and their relationship to performance on the same complex spatial task as was employed in the two previous experiments. In addition to the stress conditions used in Experiment 2 (normal and unstable conditions), we included a third condition whereby the task actually did change in terms of suddenly introducing different rules. We hypothesized that learners in the new condition would show the worst performance, overall, and this was upheld. Further, we posited that females would be more affected by the new stress condition compared to males, particularly the low-testosterone females. This also was supported by our data.

In conclusion, we are beginning to identify conditions under which we can differentially influence spatial performance. While our replication attempt failed with regard to the effects of the small discussion groups enhancing female performance, we did find that (a) there are reliable gender differences underlying spatial learning, (b) these differences are accounted for, in part, by testosterone level, (c) certain instructional conditions (like task uncertainty/instability) serve to enhance performance, overall, and (d) this performance enhancement is seen for all but the low-testosterone females.

On the basis of these findings, we continue to seek innovative ways to boost the performance of females (especially those with low testosterone levels) on similar kinds of spatial tasks. We suspect that the instructional intervention examined in Experiment 1, but carried out with video-taping and more structured discussions, would help this group, especially in conjunction with the provision of adequate practice opportunities. This would allow the low-testosterone females to become sufficiently facile with the task such that the introduction of some

task instability would then elevate their performance. Unless a person is at a certain (minimal) level of proficiency, any arousal (positive functions) associated with task changes would only hurt, not help, the learner.

In our associated paper summarizing findings from Experiments 4, 5, and 6 of this series, we attempt to examine similar issues regarding gender differences in performance and simple instructional interventions intended to reduce any differences. However, instead of employing spatial tasks, we used different criterion tasks assessing: procedural learning and quantitative skills.

References

Eysenck, M. W. (1983). Anxiety and individual differences. In R. J. Hockey (Ed.), Stress and fatigue in human performance (pp. 273-297). New York, NY: Wiley.

Gopher, D., Weil, M., & Bareket, T. (1992). The transfer of skill from a computer game trainer to actual flight. Proceedings of the Human Factors Society, 36th Annual Meeting, Atlanta, GA, Vol. 2, 1285-1290.

Jones, D. M. (1983). Noise. In R. J. Hockey (Ed.), Stress and fatigue in human performance (pp. 61-95). New York, NY: Wiley.

Kimura, D. (1992). Sex differences in the brain. Scientific American, 119-125.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. Child Development, 56, 1479-1498.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Mane, A. M., & Donchin, E. (1989). The Space Fortress game. Acta psychologica, 71, 17-22.

McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. Psychological Bulletin, 86, 889-918.

Regian, J. W. & Shute, V. J. (1993). Basic research on the pedagogy of automated instruction In D. M. Towne, T. de Jong, and H. Spada (Eds.), Simulation-based experiential learning (Series F, Vol. 122, pp. 121-132). Berlin: Springer-Verlag.

Shebilske, W. L., Regian, J. W., Arthur, W., & Jordan, J. (1992). A dyadic protocol for training complex skills. Human Factors, 34, 369-374.

Shute, V. J., Pellegrino, J. W., Hubert, L., & Reynolds, R. W. (1983). The relationship between androgen levels and human spatial abilities. Bulletin of the Psychonomic Society, 26(6), 465-468.

Spielberger, C. D. (1983). State-Trait Anxiety Inventory for Adults: Self-Evaluation Questionnaire (Form Y-1). Palo Alto, CA: Mind Garden.

Spielberger, C. D. , Gorsuch, R., & Lushened, R. (1970). The State-Trait Anxiety Inventory (STAI) Test Manual Form X. Consulting Psychologists Press.

Voyer, D., Voyer, S. & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. Psychological Bulletin, 117(2), 250-270.

Yerkes R. M. & Dodson J. D. (1908). The relation of strength of stimuli to rapidity of habit-information. Journal of Comparative Neurology and Psychology, 18, 459-482.

Appendix 1

Consent Document

VOLUNTARY INFORMED CONSENT STATEMENT
INSTRUCTIONAL INTERVENTIONS FOR REDUCTION OF GENDER DIFFERENCES IN LEARNING

1. NATURE OF RESEARCH. I hereby volunteer to participate as a test subject in the above study. I understand the purpose of this study is find effective ways to reduce experience-related gender differences in performance. Prior research has demonstrated that (a) there are certain (e.g., spatial) tasks where males often perform better than females, (b) these gender-related performance differences are probably due in part to testosterone levels and in part to experiential differences, and (c) different kinds of instructional interventions have been found to reduce the performance differences. I understand that I'll be participating in an 8-hour study conducted at Lackland Air Force Base (Armstrong Laboratory) that will measure my performance on a computerized, self-paced learning task (from 3 -7 hours, depending on my acquisition rate) as well as my testosterone level (via saliva sample). I also understand that I may be given the option to additionally participate in another 4-hour study that will be conducted approximately one month later, and involve a more "stressful" version of the learning task than the current one. Stressful versions of the task will be created by introducing a speed requirement and performing the task in a noisy environment (a tape-recording of a stock exchange during frenzied trading played via headphones). This follow-on study will examine performance as a function of stressful or non-stressful conditions and the interaction with testosterone level. Finally, I realize that to obtain valid testosterone assessments, I will have to provide some saliva (5-10 ml, or .17 oz.) in a tube that will be taken to Southwest Research Institute for a radioimmunoassay of testosterone level. This saliva/testosterone sample will be collected at the same time of day (i.e., at the outset of the study) to control for fluctuations in normal testosterone levels. Samples will be refrigerated until transferred to SwRI (Dr. W. Uhland) within 2 days.

2. PROCEDURES. As a participant, I understand that I will be answering a short questionnaire that asks about my age, sex, and any medications that I'm taking (as some medications influence testosterone levels, such as birth control pills); providing a saliva sample for the analysis of the level of testosterone; and then participating in a computer-administered learning task which measures how effectively I acquire certain skills. I'm aware that there will be at least two test administrators present to answer any questions I may have, and that my participation is voluntary. I understand that there is no known adverse effect due to any of these aforementioned procedures. I understand also that I am not to eat or smoke one hour prior to providing my saliva sample, which will be collected within the first 1/2 hour of the study.

3. RISK OR DISCOMFORT. I have been thoroughly and formally instructed on the foreseeable risks and discomforts associated with the saliva-collection procedure as well as completing the computerized learning tasks. In particular, I understand that the only potential risk or discomfort may result from staring at a computer screen for several hours which may cause some degree of eye strain. To guard against this possible discomfort, I understand that several break periods have been inserted into this 8-hour experiment (i.e., two 15-minute breaks as well as a 1-hour lunch break). Furthermore, if it becomes too problematic, I am free to terminate the experiment at any time. All of the material (i.e., responses to the questionnaire, the level of testosterone, and results on the learning task) will be kept strictly confidential; this confidentiality will be guaranteed to the extent provided by the law.

4. BENEFITS TO SUBJECTS. I understand that there is no direct benefit to me as a subject. However, one benefit of the research, in general, is that it will enable the determination of the most effective factors that can reduce (or eliminate) incoming cognitive differences resulting from hormonal discrepancies between males and females. By identifying successful instructional interventions that can *attenuate* these differences, I understand that this may serve to close the gender gap as well as open up a wider realm of new possibilities for women in all branches of the armed services.

5. ALTERNATIVE COURSE OF TREATMENT. All potential alternative procedures for measuring testosterone levels are considerably more invasive than the saliva-collection procedure. The alternatives consist of collecting blood and urine samples, both of which carry significantly more risks and/or discomforts than the selected procedure.

6. TERMINATION FROM THE STUDY. I understand that there are certain circumstances which may result in termination of my participation in this study. These possible circumstances include any computer failure that would result in a loss of data, or my lack of effort or ability to perform the learning task.

7. PARTICIPATION.

a. Records of my participation in this study may only be disclosed according to federal law, including the Federal Privacy Act, 5 U.S.C.552a, and its implementing regulations.

b. I understand that my entitlement to medical care or compensation in the event of injury is governed by federal laws and regulations, and if I desire further information I may contact the Base Legal Office, Brooks AFB at 536-3301.

c. The decision to participate in this research is completely voluntary on my part. I am participating because I want to. Dr. Wes Regian or Dr. Val Shute have adequately answered any and all questions I have about this study, my participation, and the procedures involved. I understand that the Base Legal Office, Brooks AFB will be available to answer any questions concerning procedures throughout this study. I understand that if significant new findings develop during the course of this research which may relate to my decision to continue participation, I will be informed. I further understand that I may withdraw this consent at any time and discontinue further participation in this study without prejudice to entitlements. I also understand that this experiment's medical monitor (Dr. E. Gomez) may terminate my participation in this study if he or she feels this to be in my best interests.

d. A copy of this form will be given to me.

(VOLUNTEER'S SIGNATURE & SSAN)	(DATE)
(PRINCIPAL INVESTIGATOR'S SIGNATURE & SSAN)	(DATE)
(CO-INVESTIGATOR'S SIGNATURE & SSAN)	(DATE)
(WITNESS)	(DATE)

Privacy Act of 1974 applies. DD Form 2005 filed in Clinical/Medical Records.

Appendix 2

RIA Method for Determination of Testosterone in Saliva

1.0 **SCOPE**

This procedure applies to anyone who determines the testosterone concentration from saliva.

2.0 **PURPOSE**

This procedure is to ensure safe handling of radioactive materials, check the spread of radioactive contamination, and ensure accurate determination of the testosterone concentration from saliva.

3.0 **RESPONSIBILITIES**

It is the responsibility of anyone performing the wet radiochemistry and/or counting to follow these procedures. No one is allowed to perform any radiochemical procedures until they have been cleared by the department's radiation safety officer.

4.0 **FREQUENCY**

This S.O.P. will be followed anytime a determination a testosterone concentration from saliva is to be performed.

5.0 **PROCEDURE**

5.1) Allow all tubes, reagents, and samples to warm up to ambient temperature before starting.

5.2) Label all tubes appropriately (e.g., blank, standard, sample number, etc.).

5.3) Pipette 25 λ of sample, control, or standard into the appropriate tube, or 1 ml of "clean" saliva.

5.4) Add 1 ml of physiological grade saline to all of the tubes that only contain a volume of 0.025 ml (e.g., to all tubes except those containing saliva).

5.5) Extract the testosterone by adding 2 ml of diethyl ether and shaking vigorously. Transfer the organic layer to clean, labeled RIA tube.

5.6) Add another 2 ml of diethyl ether to the remaining aqueous phase and again shake to extract any remaining testosterone. This time freeze the aqueous phase by placing the tube in dry ice, and pouring off the organic layer. Pool this with the material collected above.

5.7) Evaporate off the ether by blowing nitrogen, helium, or some other suitable gas through it.

5.8) Add 1.0 ml of the I-125 labeled testosterone into each tube and mix.

5.9) Incubate at 37° for two hours.

5.10) Aspirate off the liquid in the tubes, and collect it with the radioactive liquid waste.

5.11) Rinse the tube with two milliliters of physiological grade saline.

5.12) Cut the top off of each tube.

5.13) Place each tube in a separate liquid scintillation vial, and add 10 ml of the proper cocktail.

5.14) After the tube has set overnight, count it on the liquid scintillation counter.

6.0 **QUALITY CONTROL**

6.1) At least ten percent, but no less than one of the samples shall be run as a replicate.

7.0 **REPORTING DATA**

7.1) The count rate of the I-125 from each tube will be reported in "CPM".

7.2) The values of the blanks will be averaged, and the count rate of each calibration standard will be divided by this value.

7.3) The log of the amount of testosterone ("X" axis) is plotted against the log of the ratio of the standards' count rates to the average of the blank's count rates ("Y" axis).

7.4) For each sample the count rate is divided by the average count rate of the blanks, and using the graph generated above, the amount of testosterone present is determined. This value is then divided by the volume of saliva used to determine the testosterone concentration in pg/ml.

The Effects of Instructional Intervention and Gender on Procedural and Quantitative Tasks

(Final Report--Part B)

Valerie J. Shute

Armstrong Laboratory/HRTI

Brooks Air Force Base, Texas

Lisa A. Gawlick

Galaxy Scientific Corporation

San Antonio, Texas

The Effects of Instructional Intervention and Gender on Procedural and Quantitative Tasks

The preceding paper (Shute & Gawlick, 1996--Part A) summarized findings related to gender and instructional interventions relative to performance on a complex spatial task. Spatial tasks have consistently shown incoming gender differences, and the goals of the preceding series of experiments were to: (a) ascertain the degree of incoming gender differences, (b) seek ways to improve performance with various instructional interventions, and (c) examine the role of testosterone in relation to the remaining differences. While our initial instructional intervention (i.e., small, mixed-gender discussion groups) failed with regard to the attenuating gender effects, we did, coincidentally, find another kind of instructional manipulation that enhanced performance on the complex spatial task by creating a state of task uncertainty/instability. This technique worked well on all categories of learners except low-testosterone females. The current series of experiments continue to explore instructional interventions that improve acquisition (particularly for deficit learners), across domains.

The three experiments reported herein examine gender differences in relation to two different tasks--one teaching and assessing procedural learning, and the other, quantitative skills. While these tasks do not have the same historical reputation for yielding reliable gender differences, we are still interested in testing whether simple instructional interventions can improve performance, relative to control conditions, and specifically assessing their interaction with gender. The instructional interventions investigated in the following experiments are: (a) supplemental analogies, (b) an array of practice schedules, and (c) learner control (i.e., allowing learners to choose the amount of practice they receive for any given concept).

The first study (Experiment 4) investigates the role of employing analogies as a supplement to learning a procedural rule learning task (Procedural Phoenix), and further

examines the effects of gender in relation to performance on this task. The second two studies (Experiments 5 and 6) examine the effects of different practice opportunities on learning quantitative skills (Experiment 5), as well as the effects of treatment/practice condition and gender on the retention of quantitative knowledge and skills following a 6-month period (Experiment 6). In those experiments, we employ a statistics tutor (Stat Lady) teaching descriptive statistics.

Experiment 4

Research involving analogies has shown that knowledge and skill acquisition can be facilitated through the use of effective analogies (e.g., Gentner & Gentner, 1983; Glynn, 1991). Effective analogies should be immediately recognizable by the learners, so they can map this new knowledge onto pre-existing concepts, and increase skill acquisition of the new knowledge and skill. Thus, attempts at improving instruction should capitalize on these aspects of skill acquisition.

How do analogies facilitate learning? Most current learning theories maintain that new knowledge and skills are acquired by building on what is already known. Analogies are useful because they are based on well-known concepts (e.g., balancing a seesaw as an analog for finding the arithmetic Mean). Thus, learners acquire knowledge and skills by actively interweaving new knowledge with existing knowledge (e.g., Barlett, 1932; Collins, Brown, & Newman, 1989; Piaget, 1954).

The learning theory behind ACT-R (Anderson, 1993) asserts that all knowledge begins in a declarative form, and productions (used to carry out procedures) develop through an analogy process. Taking this further, the learning theory behind SMART (Shute, 1995) proceeds from symbolic knowledge acquisition (i.e., bits of knowledge and skill at the lowest level) to

procedural skill acquisition (i.e., the application of that knowledge), to finally, conceptual understanding. This third stage (conceptual understanding) is believed to be best instructed and remediated (and hence learned) via analogies (see Shute & Catrambone, 1996).

The primary research questions investigated in this experiment relate to testing: (a) the effects of supplementing instruction with analogies, (b) gender differences in relation to declarative and procedural knowledge outcomes, and (c) performance differences due to a treatment by gender interaction.

The criterion task used in the following experiment was a desk-top flight simulator (Phoenix). Learners were trained and tested on the declarative knowledge and skills of the task (e.g., understanding basic concepts, interpreting gauges, maneuvering the plane through gates) in either an analogy or control condition. Our hypothesis concerning the analogy condition was that, generally, learners in this condition would perform better than a control condition (receiving only standard instructions). The second phase of the experiment involved training on procedural rules in a slalom task (i.e., flying planes through appropriate gates according to given rule gates). We wanted to examine the data for possible gender differences, and hypothesized that males would perform better than females on this task its additional spatial component. Additionally, if females were having difficulty acquiring the declarative and procedural skills, we expected them to benefit from supplemental analogy instruction moreso than the males.

Method

Participants. A total of 69 individuals participated in this experiment (47 males and 22 females). All participants were recruited through local temporary agencies in San Antonio, Texas and paid approximately $5.00 per hour for their involvement. Participants ranged in age from 18 to 30 years (M = 22.0, SD = 2.5), and had a high school diploma or GED, but had not completed

a four-year college degree. Screening was conducted to eliminate those persons with prior exposure to the task (Phoenix).

Equipment. The experiment was conducted in the USAF Armstrong Laboratory located at Lackland Air Force Base, Texas. The laboratory consists of 30 Compaq 486/33L computers with NEC/Multisync SVGA monitors. Learners used standard keyboards and CH Products FlightStick joysticks.

Task. Phoenix is a Desktop Flight simulator (with an "out-of-the-cockpit" view) used to provide learners with experience in either shooting targets (Strike Task) or flying through an airborne slalom course (Slalom Task). We used the Slalom Task in this experiment, where learners were first instructed on simple flight procedures (e.g., adjusting the roll or pitch of their plane), and then on how to maneuver their plane through gates suspended in the airspace. The on-line introduction to the task familiarized learners with the orientation of the Heads Up Display (i.e., interpreting the speed, heading, pitch, and altitude dials) and basic maneuvering skills (e.g., rolling the plane to a specified setting). The Slalom Task specifically required learners to interpret a Rule Gate according to the presented rule, and then maneuver their plane accordingly. For more information on this task, see Goettl & Shute (in press).

Procedure. In this experiment, we incorporated analogies into the instruction of declarative knowledge and skills associated with flying the plane in the simulated environment (i.e., pitch, heading, altitude, speed, thrust, and roll). Following the generic, on-line introduction to the task, learners in both conditions (analogy and control) received verbal instructions explaining these basic flight skills. However, learners assigned to the analogy condition received their explanation as analogies, illustrating each concept.

For learners in the control condition, the concept of *roll* was explained as the position of the plane's wings relative to the ground. Rolling the plane to the left causes the left wing to move down and the right wing to move up, as the plane moves to the left. On the other hand, for learners in the analogy condition, the concept of *roll* was compared to placing tennis balls on a tray and rolling them left and right. As the balls roll to the right, the left side of the tray rises as the right side lowers. Learners in both conditions were given these verbal instructions in separate rooms and then returned to their testing stations.

This experiment was a 2×2 factorial design: treatment *condition* (analogy vs. control) $\times$ *gender* (male vs. female). Individuals were trained on the Phoenix task in groups of approximately 15 individuals over an 8-hour period. The on-line introduction familiarized learners with the Phoenix task, and was administered first. Next, participants were randomly assigned to either the analogy or control condition, and verbally instructed on the task. Then all participants were given training across 10 trials, followed by a paper and pencil posttest assessing declarative knowledge of basic flight concepts (e.g., heading, pitch, altitude). Learners were then instructed and tested on the procedural rules for the slalom task.

Declarative Knowledge and Skills. Following familiarization with the simulator (i.e., the on-line introduction), learners received training (over 10 trials) on 4 specific tasks: (a) *Unroll* (the plane is presented in a rolled position, and learners must unroll the plane to 0 degrees within 20 seconds), (b) *Roll* (the plane is presented in a level position, and learners must roll the plane a specified number of degrees within 20 seconds), (c) *Heading* (the plane is initially placed at a specified heading, such as 270 degrees, and learners must adjust the plane to another specified heading within 3 minutes), and (d) *Gates* (learners have 30 seconds to maneuver their plane

through three-dimensional gates within their visual field). A paper and pencil test was created to assess these basic concepts (e.g., speed, heading, thrust, pitch, roll).

Procedural Rules. Four procedural rules were created, differing on the number of IF and THEN clauses making up the rule (i.e., singular vs. multiple IF and THEN statements). The simplest rule had one IF condition and one THEN condition (e.g., IF there are three targets across the top of the rule gate, THEN fly through the GREEN gate; ELSE fly through the BLUE gate), while the most complex rule had two IF conditions and two THEN conditions (e.g., IF there are three targets on the left AND all are different colors, THEN first fly through the BLUE gate AND NEXT fly through the YELLOW gate; ELSE fly through the YELLOW gate). Thus, the rules were distinguished as simple vs. complex rules, as well as one-gate vs. two-gate rules (i.e., simple/one-gate, complex/one-gate, simple/two-gates, complex/two-gates).

Each rule was instructed/tested over four trials. Before each trial, learners were presented with a rule in the text window at the bottom of the screen. Once the trial started (by pressing the space bar), the Rule Gate was displayed directly in front of the plane's starting position. Learners had 180 seconds to interpret the Rule Gate according to the given rule and take action (i.e., maneuver their plane through the gate(s) presented on the left or right side of the screen). The thrust was fixed at 20% of maximum, thus learners had to interpret the Rule Gate quickly, before the Rule Gate was no longer in their field of vision. Learners only had control over their heading and roll, which was adequate for maneuvering their plane to the appropriate gate(s).

The Rule Gates themselves were made up of several three-dimensional octahedrons, suspended in the simulated environment. The octahedrons were of varying colors and arranged in a 3×3 matrix (although not all of the cells were filled for a specific Rule Gate). Each trial ended

when either (a) 180 seconds passed, or (b) the plane reached a certain point in space behind the last gate in the trial. When the learner was ready to continue, a new trial commenced.

Results

We examined declarative knowledge and skill acquisition by separately analyzing performance data related to: (a) the training trials, and (b) the paper and pencil tests assessing declarative knowledge. For the *training trials*, we computed the average percent correct across trials for each of the four tasks: unroll, roll, heading, and gates. A MANOVA was computed on the training trials using the four percent correct scores as our dependent measures, with *condition* (analogy vs. control) and *gender* (male vs. female) as our between-subjects variables. No significant main effects were found for *condition*: $F(4, 59) = 0.61$, $p = .66$, or *gender*: $F(4, 59) = 2.27$, $p = .07$, nor was the interaction significant: $F(4, 59) = 0.80$, $p = .53$. Thus, during training, learners did not differ on declarative knowledge and skill acquisition as a function of treatment condition, gender, or the interaction of the two variables.

We then tested for differences in performance on the paper and pencil tests by computing an ANOVA on overall percent correct scores, using *condition* and *gender* as our between-subjects variables. Again, no significant main effects of these variables were found (Fs < 1). However, the *condition* × *gender* interaction was significant: $F(1, 25) = 4.37$, $p = .05$. Males performed significantly better on this test having learned from the analogy condition compared to the control condition. Conversely, females performed significantly better in the control condition rather than the analogy condition. These data are shown in Figure 1.

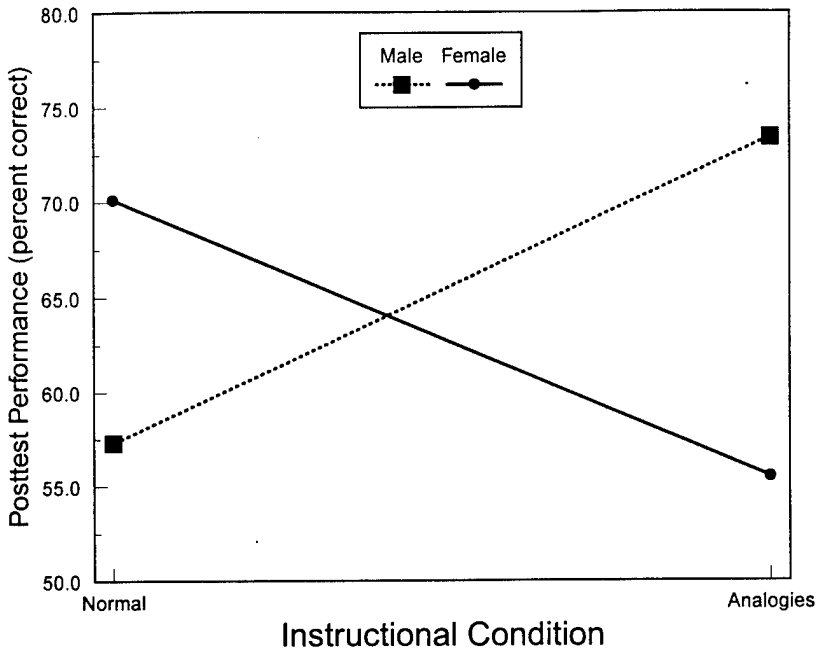


Figure 1. Gender by condition interaction on declarative knowledge tests.

To test for acquisition differences across the procedural-rule trials (4 tests for each of the 4 rules), we computed a MANOVA on the accuracy and latency data (2 separate analyses), using gender and condition as our between-subjects variables. Results showed no significant main effects or interactions on these variables (all Wilks' exact Fs < 1). Thus, the analogy instruction appeared to have no effect on procedural skill acquisition, learners did not differ in performance due to gender, nor was there any condition by gender interaction.

Discussion

Results did not show any significant main effects of gender on declarative and procedural knowledge acquisition. Nor did we find any main effects with regard to our instructional intervention (i.e., supplemental analogies). Further, our hypothesis concerning the condition × gender interaction was not supported. That is, while the condition by gender interaction was significant, it turned out to be exactly the opposite of what we had predicted--females actually benefited more from the standard, not the analogies instruction; males performed best in the analogy condition (compared to the control).

One possible explanation for this finding involves the examination of the analogies themselves. For an analogy to have value and be effective, the base analog (e.g. a seesaw) must be better known by the learner than the target analog (e.g., the arithmetic Mean). If the base analog is relatively easily understood, then the teacher or tutor must be concerned with making sure the mapping is straightforward, robust, and reliable. In our study, the supplemental instructions used two analogies for each of the five concepts. In each case, the first analogy used a car to illustrate the relevant concept while the second analogy used a computer (for two illustrations), a balloon on a string, a person's gaze, and tennis balls on a tray. Although originally we believed that these analogies would be well-known, it may be that the females in our sample were less familiar with cars and computers as they would have been with other illustrations.

Another plausible explanation for our obtained interaction may be due to the mediating effect of some cognitive ability. For instance, males typically have better spatial skills than females. Because our analogies were presented verbally, an extra step was required to translate the analogy into a spatial image (e.g., tennis balls moving on a tray as a function of the tray being

tilted). Thus, individuals with better spatial visualization skills (e.g., males) would benefit more from the analogy condition compared to learners with lower skills. We suspect that the analogies would have been more universally effective had they been represented dynamically on the computer screen. In that case, the "extra step" would be removed, as well as the hypothesized mediating effects of spatial skill (and hence, the male advantage). While we did not get a chance to test this hypothesis directly in the current study, we plan to collect cognitive data in future studies to examine individual differences in relation to the effectiveness of analogies.

The next series of research questions address the effects of practice on knowledge and skill acquisition; specifically, the influence of different practice opportunities on knowledge and skill acquisition in a quantitative domain (descriptive statistics). We report findings from two studies where the treatment conditions differ along a continuum of minimal to extended practice. We also report findings related to the effects of a new condition where learners control the amount of their practice (i.e., select how many problems they wish to solve, per concept). Thus, within this domain, we examine the effects of gender, amount of practice, and learner control on learning outcome, efficiency, and retention.

Experiment 5

Learning represents a change in a person that occurs at a particular time as a function of experience or practice. Because it's not directly observable, learning must be inferred from performance on a test, where the retention interval may be immediate or delayed. Numerous studies have been conducted suggesting that the relationship between practice, acquisition, and retention is not quite as straightforward as the "practice makes perfect" premise suggests. That is, the literature is full of findings that show how, relative to a "standard" practice condition, certain

acquisition conditions that slow the rate of improvement, or decrease performance during practice, still yield enhanced post-training performance (see Schmidt & Bjork, 1992).

The literature on learner control is even less definitive. Computerized learning environments can be characterized by the amount of learner control supported during the learning process. This dimension can be viewed as a continuum ranging from minimal (e.g., rote or didactic environments) to almost complete learner control (e.g., discovery environments). Two opposing perspectives address the issue of the best learning environment to build in intelligent instructional software. One approach is to develop an environment which provides the learner freedom to explore and learn (e.g., Collins & Brown, 1988; Kinzie, Sullivan, & Berdel, 1988; Shute, Glaser, & Raghavan, 1989). The other approach argues that it is more efficacious to develop directive learning environments (e.g., Corbett and Anderson, 1989; Sleeman, Kelly, Martinak, Ward, & Moore, 1989). Actually, this disparity may be resolved by, instead of looking for main effects of learning environment, one should additionally consider learner characteristics with the goal of identifying optimal learning environments for specific kinds of persons. Do males and females differ, in general, with regard to optimal learning environment?

The following experiment investigated the effects of various practice opportunities on quantitative skill acquisition and outcome. The first goal was to replicate findings from a previous study that varied the number of problems solved per problem set in each condition (Shute & Gawlick, 1995), in an attempt to generalize findings across domains (i.e., from flight engineering knowledge and skills to introductory statistics). Results from the original study showed that learners receiving an abbreviated practice schedule (few problems) completed the curriculum in half the time it took those learning from an extended schedule (more problems), but at the expense of greater errors and latencies during problem solution within the tutor.

Despite the acquisition differences, learners in the different practice conditions performed the same across all learning outcome measures. In the current experiment, we replicated the same practice conditions used in the original Shute & Gawlick (1995) study: Abbreviated (AA), Extended (EE), and two mixed conditions (AE and EA) that received half of their instruction under one condition and the other half under the other condition.

The second goal was to examine the effects of including a new practice condition, Learner Choice (LC), which allows learners to select the number of problems to solve rather than solving a fixed number of problems. By including this new treatment condition, we can test the effects on these same learning parameters when learners are in control of their practice environments. Do individuals, in general, have the necessary metacognitive skills to know when additional help is needed, or when they've had enough practice?

The third goal of this experiment was to examine the data for possible gender differences. Specifically, we're interested in determining whether the LC condition is an intervention that motivates learners to become more active in the learning process, thus resulting in higher outcome scores compared to other conditions. Do males and females differ in terms of relative gains from the LC condition?

Hypotheses.

1. *Skill Acquisition.* Based on findings from the original study, learners assigned to the more limited practice conditions (AA, AE) were expected to exhibit more errors during learning compared to participants learning from the more extended conditions (EE, EA) given fewer practice opportunities in which to apply newly-developing knowledge and skills. We further expected learners, assigned to the LC condition, to perform about average during skill acquisition, making a moderate number of errors compared to the other conditions. This was

based on the belief that participants in this condition would elect to solve a range of problems (from very few to very many) based on individual differences in general aptitude, metacognitive skills, and personality traits. The end result was expected to balance out at some middle level of performance.

2. *Learning Outcome*. We predicted no differences on the posttest measure among groups, given findings from the original study. However, if there *were* any differences, we expected learners in the most extended conditions (EE and EA) to perform better on the posttest compared to learners in the abbreviated conditions (AA and AE) given they would have had significantly more practice opportunities. With regard to the LC condition, we speculated that these individuals would perform at some intermediate level in terms of outcome performance given more variability in the number of problems they chose to solve. That is, to be effective within this condition, learners must be aware of cognitive strengths or weaknesses--choosing the amount of practice most appropriate for their needs, per problem set. However, many people are not so aware, and consequently may solve too few, or too many problems. Further, learners who tend to be overly confident in their abilities (or slothful, in general) may not elect to solve additional problems, when needed. Thus, learners, randomly assigned to this condition, would represent a range of these proclivities, and consequently "balance out" with regard to the number of problems they elect to solve (hence, the intermediate outcome postulate).

3. *Learning Efficiency*. The time taken to complete the tutor should be a direct function of practice condition. Thus, learners in the most abbreviated conditions (AA, AE) would take the least amount of time to complete the curriculum given fewer problems to solve, and learners in the most extended conditions (EE, EA) would take the most amount of time. Learners in the LC condition were expected to take an intermediate amount of time. Again, this was based on our

belief that learners are often not cognizant of their cognitive strengths and weaknesses, nor are many of them sufficiently motivated to continue practicing until a skill is mastered. Specifically, we predicted the following ordering of conditions in terms of tutor-completion times: AA < AE < LC < EA < EE.

4. *Gender by Condition Interaction.* We posited that females may demonstrate better outcome scores within the more extended practice environments (compared to other conditions) while males may show optimal performance within the independent LC condition. In partial support of this premise, Newcombe (1982) reported that males report themselves as more active, independent, persistent, and interested in math and science compared to females (p. 237). Thus, the LC condition may be most appropriate for males, and the extended conditions more suitable for the female participants.

Method

Participants. A total of 380 individuals participated in this experiment, obtained from local temporary employment agencies. The age range of the sample was between 18-30 years (M = 22.0, SD = 3.5), and all had a high school diploma or equivalent. Overall, 66% of the sample was male, and no one had any prior exposure to statistics courses. Participants were paid for taking part in the study and informed that they needed to return in 6 months for phase 2--retention testing.

Materials. The first module of the Stat Lady Descriptive Statistics series (DS-1, Shute & Gluck, 1994) was used as the complex learning task in the experiments described in this paper (for more on this module, see Shute, 1995). The curriculum was decomposed into low-level curriculum elements (CEs), representing units of instruction that vary in grain size, from very low-level bits of knowledge and skill, to more global units. In this study, participants received 77

CEs, arranged from simple to more complex concepts and skills, spread across five main problem sets or topics: (a) frequency distributions, (b) proportions and percentages, (c) grouped frequency distributions, (d) cumulative frequency distributions, and (e) plotting. Each CE (or small groups of related CEs) was instructed in an interactive manner, then assessed in terms of degree of mastery. In this study, the number of problems that learners solved was solely a function of assigned condition. All learners had to answer a number of CE-related questions before moving on to subsequent CEs. If a learner gave an incorrect answer to any problem, Stat Lady intervened with error-specific feedback.

Duplicate items were created to assess each of the 77 CEs. This resulted in two parallel forms (A and B) of a test that was administered on-line, before and after the tutor. For more details on these tests, see Shute (1995).

Design and Procedure. The design was a 5 × 2 factorial: *condition* (AA, AE, EA, EE, LC) × *gender* (male, female). In the original study, participants were either switched to a new practice condition (e.g., A-->E), or remained in the same one (e.g., E-->E) about 3/5 of the way through the tutor. Similarly, in this study, learners (not in the LC condition) were either switched to a new condition or remained in the same one, after the 3rd (of 5) problem sets. The five practice conditions were: (a) AA, (b) AE, (c) EA, (d) EE, and (e) LC.

Participants were tested in groups of about 20, and randomly assigned to a condition. For each of the two sections of the tutor, participants in the abbreviated (A) condition solved 1 problem per problem set, and in extended (E) condition, they solved 4 problems per set (maintaining the 4:1 ratio established in the original study). The total number of problems presented, per condition, were: AA (5), AE (11), EA (14), EE (20), and LC (variable, between 5 -

20). The sample sizes per condition in this study were: AA (n=86), AE (n=60), EA (n=58), EE (n=88), and LC (n=88).

On-line demographic questionnaires and pretests were administered to all participants. After completing both of these activities, they proceeded to learn from the tutor which took, on average, about 5 hr to complete. Finally, all participants were administered an on-line posttest assessing the full range of knowledge and skills acquired from the tutor.

Results

Prior to making comparisons between practice conditions, we needed to insure that learners within each condition were demographically comparable. Several one-way ANOVAs were computed on age, gender, number of years of education, and computer experience, by condition. None of these variables showed significant differences across the five groups.

Skill Acquisition. Does treatment condition, gender, or the interaction of these two variables affect acquisition accuracy? We examined this issue first by comparing the *number of errors* made (i.e., frequencies of negative feedback, averaged across all CEs). There was a significant main effect due to *condition*: $F(4, 363) = 7.46$, $p < .001$. The order of average errors per CE across conditions was: LC (1.67) < AA (1.70) < AE (2.41) < EA (2.47) < EE (2.96). There was also a significant main effect of *gender*: $F(1, 363) = 4.26$, $p = .04$ where males committed fewer errors ($M = 2.08$) relative to females ($M = 2.50$). The condition by gender interaction was not significant.

The number of errors one makes is related to the number of questions received, which also differed by condition: $F(4, 368) = 781.38$, $p < .001$, but not by gender or the interaction between condition and gender. Thus, to test for differences in *acquisition accuracy*, we created an error-rate variable--the number of errors divided by the number of questions, averaged across

CEs. For this index, values closer to 1.0 denote average performance--an equal ratio of errors to questions; values *less than* 1.0 denote more accurate performance because fewer mistakes are being made relative to the number of questions answered, and values *greater than* 1.0 denote more inaccurate performance as more errors are committed relative to questions received.

For our sample, this value ranged from 0.40 to 2.09 and was significantly different among conditions: $F(4, 363) = 17.55$, $p < .01$. There was no main effect due to gender, nor was the interaction between condition and gender significant. The order of this variable by condition was: EE < EA = AE < LC < AA. Thus, as with the previous study, the EE learners' *error rate* was the lowest among conditions--they made fewer mistakes relative to their greater number of questions. Learners in the AA condition showed the highest *error rate*--they tended to commit more mistakes on relatively fewer questions. Learners in the LC condition showed error rates that were about midway between the AA and the EE conditions. See Figure 2.

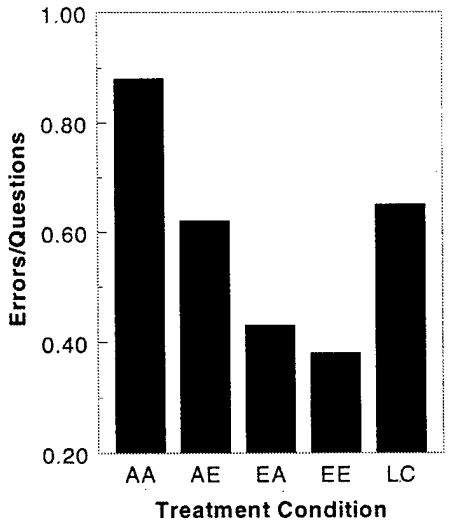


Figure 2. Error rate data across five treatment conditions

Learning Outcome. We first examined the pretest data to insure our two between-subjects variables (condition and gender) were comparable in terms of prior knowledge and skills related

to statistics. We computed an ANOVA on pretest score, by condition and gender, and found significant differences among *conditions*, $F(4, 370) = 2.39$, $p = .05$. Specifically, the EE group (by chance) began with the highest pretest Mean (greatest incoming knowledge), the AA with the lowest, and the LC participants, in between. Thus, we used pretest score as a covariate in subsequent analyses. There was no main effect of gender or a condition by gender interaction in relation to pretest data.

An ANOVA was computed on the posttest data (Means adjusted for pretest score) by condition and gender. There was no main effect due to *condition*, although the results were marginally significant: $F(4, 370) = 2.36$, $p = .052$. There was also no main effect of *gender:* $F(1, 370) = 1.95$, $p = .16$, or *condition × gender*: $F(4, 370) = 1.16$, $p = .33$. The order of posttest scores by condition was: AA (68.3) < AE (71.4) < LC (71.9) < EA (74.5) ≈ EE (74.6).

Learning Time. We decomposed the total tutor time variable into two parts--instruction and problem-solving time. Instruction time should vary in relation to one's facility in acquiring and understanding the new material, while problem-solving time should vary in relation to condition. Three ANOVAs were computed on instruction time, problem-solving time, and total time required to complete the tutor (i.e., instruction time + problem-solving time) by condition and gender. All three variables showed significant differences due to condition. The order of total time by condition was: LC < AA < AE = EA < EE. See Table 1 for Mean times and F values.

Table 1. Instruction, Problem-Solving, and Overall Tutor Time (hrs) by Condition

TIME	**AA** ($n = 86$)	**AE** ($n = 60$)	**EA** ($n = 58$)	**EE** ($n = 88$)	**LC** ($n = 88$)	**F** ($df = 4, 375$)	**p**
Instruction	2.05	1.88	1.67	1.52	1.78	8.78	< .001
Prob-Slvg	1.98	2.96	3.16	3.96	2.13	40.38	< .001
Total	***4.03***	***4.84***	***4.85***	***5.47***	***3.91***	***12.74***	***< .001***

Notes: A = Abbreviated, E = Extended practice condition. LC = Learner's Choice practice condition.

We also found a significant main effect due to gender on these three variables. In all cases, males required significantly less time compared to females to complete the tutor. The results from the total time variable were: $\underline{F}(1, 370) = 6.63$, $\underline{p} = .01$, Male $\underline{M} = 265.9$ and Female $\underline{M} = 293.1$ (in minutes). The three associated gender by condition interactions on these time variables were not significant.

The final variable to be examined in this section combined outcome score (i.e., adjusted posttest data) and tutor-completion time to yield an *outcome-efficiency* index (i.e., posttest/time). The interpretation of this variable is that larger values reflect greater efficiency (i.e., higher learning outcome scores relative to time spent on the tutor); lower values indicate less efficient learning. We computed an ANOVA on this ratio by condition and gender. Results showed a significant main effect due to *condition*: $\underline{F}(4, 370) = 6.23$, $\underline{p} < .001$. The ordering of the efficiency index, by condition, was: EE < EA < AE < AA < LC. As can be seen in Figure 3, LC learners showed superior learning efficiency relative to the other conditions.

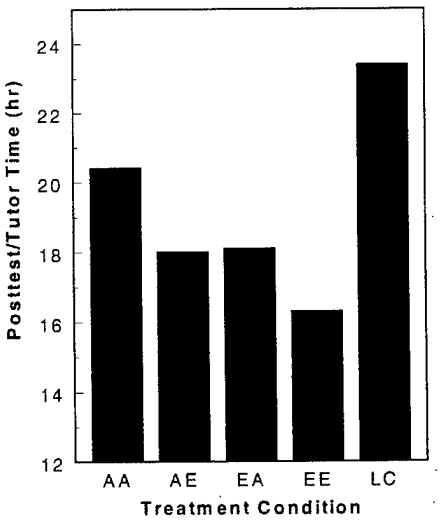


Figure 3. Outcome efficiency indices across five treatment conditions

We also found a significant main effect due to *gender*: $\underline{F}(1, 370) = 5.34$, $\underline{p} = .02$, where males ($\underline{M} = 20.33$) showed a greater efficiency index compared to females ($\underline{M} = 17.62$). The interaction was not significant.

Discussion

The first purpose of this study was to replicate rather unexpected findings from our original study (Shute & Gawlick, 1995). Specifically, we replicated the following: (a) Reduced practice opportunities result in greater errors produced during skill acquisition, but (b) despite these acquisition differences, outcome performances across all five conditions were not statistically different (even in this new domain), and (c) learners in the abbreviated condition(s) completed the tutor significantly faster than learners in the more extended conditions. When we viewed tutor-time data separated into its component parts (instruction and problem-solving time), we saw, predictably, that problem-solving time increased as a function of practice condition (AA < AE < EA < EE). However, instruction time showed a *reversal* of this ordering: EE < EA < AE < AA. This suggests that abbreviated learners may have been attempting to compensate for their sparse practice environments by spending relatively more time reviewing the instructional sections of the tutor.

A second goal of this study was to examine the effects of learner control on these same dependent measures (i.e., acquisition, outcome, and efficiency). While the LC learners did show an intermediate level of acquisition accuracy (falling in between the AA and EE groups), surprisingly, the LC learners completed the tutor faster than any other condition. Moreover, when we computed and tested an outcome-efficiency index by condition, results showed the LC learners greatly surpassed the other groups (see Figure 3). This suggests that the increased control

in one's learning environment may have motivated learners to remain more active in the learning process, resulting in greater outcome efficiency.

The final goals of this study were to determine whether any gender differences existed in this domain, and if practice conditions were differentially effective for males and females (i.e., testing the gender by condition interaction). None of the variables examined (e.g., errors during acquisition, tutor time, outcome scores) showed any interactions between gender and learning condition. Both males and females displayed comparable levels of performance relative to practice condition. However, main effects of gender surfaced in the following: (a) the relative number of errors made during learning, (b) instructional, problem-solving, and total tutor time, and (c) outcome efficiency (i.e., posttest/ total tutor time). In all cases, males performed better than females--they made fewer errors during learning, spent less time in instruction and solving problems, and were more efficient learners compared to the females.

The final experiment examines possible effects of original practice condition and gender in terms of retention. Participants in Experiment 5 were called back to participate in a retention study six months following original learning from Stat Lady.

Experiment 6

The following experiment examines the generalizability of another unexpected finding reported by Shute & Gawlick (1995) regarding greater retention for learners assigned to mixed practice conditions compared to homogenous ones. That is, after two years, individuals who had originally learned from a variable-practice schedule (i.e., mixed-group condition: AE & EA) showed significantly greater retention compared to either of the extreme/stable practice groups (AA & EE). These findings were unexpected, and the current experiment attempts to replicate the effects of practice schedule on retention. Additionally, we examine the effects in relation to

our new LC condition, where learners were in control of the amount of practice received per problem set. In particular, we were interested in testing for retention differences as a function of original practice condition, gender, and their interaction.

This experiment is still in progress--we currently have data from only 69 (of the original 380) participants. Because of the incomplete status of this phase of the study, the following should be viewed as preliminary analyses and tentative conclusions.

Hypotheses

Based on our prior findings, we expected learners in the mixed practice conditions (AE & EA) to show greater retention of the material compared to learners in the homogeneous conditions (AA & EE) following a 6-month lag between original and retention testing. Additionally, we hypothesized that learners who had been assigned to the LC condition would show average, to above-average levels of retention based on a fairly typical finding in literature which suggests that increased control over one's environment renders the learning experience more enjoyable, particularly for high-ability learners (e.g., Shute & Gawlick-Grendell, 1994; Swanson, 1990). New knowledge and skills may thus be more memorable.

Given no main effect of gender on adjusted posttest score (from Experiment 5), we did not expect to see a main effect on retention in the follow-on study (holding posttest score from Experiment 5 constant). Regarding the interaction, we wanted to test for gender differences in retention, especially for those having learned within the LC condition. Our specific gender × condition interaction posited that males would remember more had they learned in the LC condition compared to females in the same condition, while females were expected to show greater retention having originally learned within more extended conditions. This hypothesis was motivated by Shute & Gluck (in press) who reported that males showed significantly more

independent/exploratory behaviors than females when learning from an on-line instructional system, and this tendency would be most suitable for the LC condition, resulting in increased retention.

Method

Participants in Experiment 5 were asked to return 6 months after learning from Stat Lady to take part in the follow-up portion of this study. Currently, 18% of the original sample has returned (n = 69). Not all of the original participants are able to return, but we expect at least 1/3 of the sample to return. To motivate their return, we offer a monetary bonus. The average lag between original and retention testing = 26.1 weeks (SD = 1.9 weeks). The current distribution of the returning sample's original condition is AA (n=13), AE (n=10), EA (n=13), EE (n=11), and LC (n=22).

Testing is being conducted in groups of 1 to 5 persons, over one day. Prior to taking the first retention test, test administrators brief each group on the importance of trying to remember as much as they can from their original session. After the first test has been completed, participants are given a 30-minute break, followed by the second retention test.[1] At the conclusion of the second test, all returning participants are administered an on-line battery of cognitive ability tests assessing working memory capacity, information processing speed, inductive reasoning skill and fact learning ability in the quantitative domain. This battery requires, on average, about one hour to complete.

Results

[1] Both retention tests consist of items that were developed to be isomorphic to items comprising the tests employed in Experiment 5.

Prior to making any comparisons among conditions on the retention measures, we needed to insure that the subset of returning learners were comparable to the original sample in Experiment 5 (overall, and for each practice condition). We computed one-way ANOVAs on demographic measures (age, gender, education, computer experience), by experiment. None of these measures were significantly different. We also compared returning to original participants' data on Experiment 5 posttest scores (adjusted for pretest). Scores from the returning sample ($\underline{M}$ = 72.9, $\underline{N}$ = 69) did not differ significantly from the original sample ($\underline{M}$ = 72.0, $\underline{N}$ = 380), $\underline{t}(447)$ = -0.48, $\underline{p}$ = .64 on this measure.

Next, we computed a factor analysis (principal components analysis) on the cognitive ability test data (percent correct scores). This resulted in the extraction of a single factor: general aptitude. The percentage of variance accounted for by this factor was 62.3%, with $\underline{M}$ = 0, and $\underline{SD}$ = 1. Factor scores were saved for each person and used as a covariate in subsequent analyses.

We then combined data from individuals originally learning from the AE & EA conditions because: (a) their acquisition, outcome, and efficiency data from Experiment 5 were not significantly different to warrant their separation, (b) this increases the power of the upcoming analyses, and (c) this same procedure was followed in the original Shute & Gawlick (1995) study. Furthermore, we combined the two retention test scores into an average score, resulting in a more reliable measure.

To test our hypothesis concerning condition and gender effects on retention, we computed a repeated-measures ANOVA with *retention* as the within-subject variable (i.e., the difference between the adjusted posttest scores from Experiment 5 and average retention scores from Experiment 6) with *condition* (AA, AE/EA, EE, LC) and *gender* (male, female) as the between-subjects variables. We included the *aptitude* factor score as a covariate in the equation to control

for any differences in aptitude that may account for any obtained main effects or interaction. Results from the ANOVA showed no main effect on retention due to original practice *condition* ($\underline{F} < 1$), no main effect of *gender* ($\underline{F} < 1$), but a significant *condition* × *gender* interaction: $F(3, 60) = 4.86$, $p = 0.004$. This interaction is depicted in Figure 4.

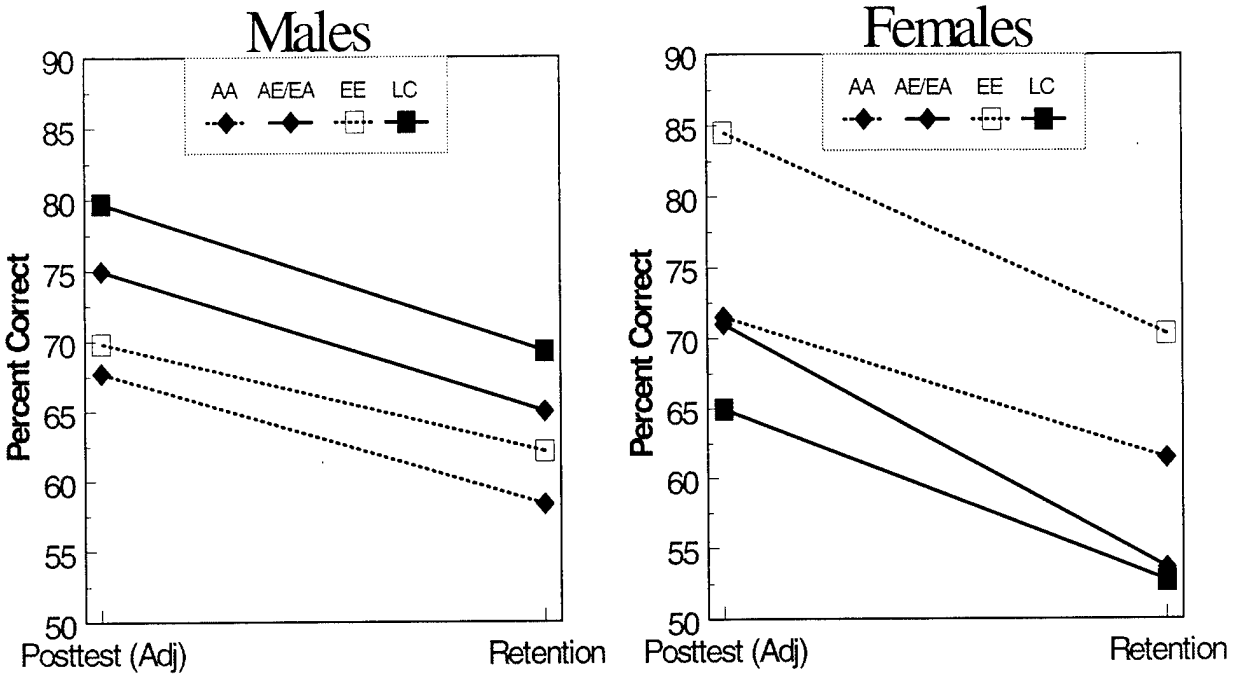


Figure 4. Condition by gender interaction on retention

This figure clearly shows that for males, learning from the LC condition represented the superior learning environment, followed by the mixed-practice conditions. In contrast, the LC and mixed conditions (i.e., those with variable practice opportunities) provided the worst environments for learning and retention for the female participants, while the most extended condition (EE) comprised the optimal environment, followed by the other stable condition: AA.

Discussion

In Experiment 6, we examined the effects of practice condition and gender on retention, and attempted to replicate findings from a prior retention study (Shute & Gawlick, 1995). In the

original study, the groups showing the greatest retention of flight engineering knowledge and skills were the mixed-practice conditions (AE and EA). Following 6 months between original instruction and retention testing in the present study, using a different domain, we found no significant main effects of original practice condition on retention, thus we failed to replicate the finding of the mixed-practice condition's advantage on retention. However, we did find an interesting condition by gender interaction showing that males performed better, and remembered more, having learned under more variable practice conditions (LC and AE/EA). Females performed optimally under the more stable practice conditions (EE and AA).

The reasons for this finding may be due to (a) gender differences in relation to the acquisition of knowledge and skills in this domain (Experiment 5), and (b) findings reported in the preceding paper (Shute & Gawlick, 1996--Part A). That is, we reported gender differences in acquiring quantitative knowledge and skills in Experiment 5 (this paper), with a male advantage (i.e., males performed better than females in terms of acquisition accuracy and learning efficiency). Findings from Experiments 2 and 3 (Shute & Gawlick, 1996--Part A), showed a significant interaction within a different domain (complex spatial task) involving task instability by testosterone level--males (and high-testosterone females) showed enhanced learning under unstable task conditions (relative to a control) while low-testosterone females showed worse performance in the unstable condition, relative to a control group.

Our explanation of this finding was that the unstable task provoked more cognitive arousal and hence resulted in better performance, but only when some minimal level of proficiency has been attained. The current interaction (Experiment 6) suggests that, again, males may be more cognitively aroused under the variable (unstable) conditions, resulting in superior performance relative to the homogeneous (stable) conditions. Furthermore, findings from

Experiment 5 showed that males initially acquired the quantitative subject matter more efficiently compared to females. Females, in general, appeared to have more difficulty acquiring the new concepts and skills, thus we suspect that they found the stable, homogeneous conditions considerably less distracting (i.e., more supportive and predictable) compared to the variable ones. So, cognitive arousal can only serve to facilitate learning if one has attained a minimal level of proficiency; otherwise, introducing variability is expected to harm, not enhance, subsequent performance.

Summary and Conclusions

The most interesting finding from Experiment 4 (declarative and procedural rule acquisition) relates to the obtained condition × gender interaction. This finding, while opposite of what we had originally predicted, indicates that females score higher on the outcome measures when learning from the standard condition, not the analogies one. Males perform better in the analogy condition rather than the control. We offered two viable reasons for the interaction: (a) males were more familiar with the base analog (cars, computers) than females, thus the mapping to the target analog (pitch, roll) was much easier for them, and (b) males, with greater spatial skill, may have profited more than females from the analogy condition as they were able to easily translate the analogy into a functional spatial image. Both of these reasons comprise empirical questions that we plan to research in future studies. For instance, one upcoming study (Shute & Catrambone, 1996) will explicitly test the efficacy of analogies on conceptual understanding (in statistics) when they are presented on the computer screen, and manipulable by the learners. We believe that this instructional manipulation should substantially shrink (or eliminate) the gender by condition interaction shown in Experiment 4.

The findings from Experiments 5 and 6, collectively, suggest that the design of automated instructional systems may be enhanced, and learning efficiency improved, by providing for greater student control during learning. Findings from Experiment 5 showed that during learning, participants in the LC condition ended up with the highest "efficiency indices" compared to all other conditions (i.e., high outcome scores relative to acquisition time). Learners who were given control of their practice opportunities performed no differently on the outcome measure compared to those having more extensive practice opportunities, and required significantly less time to attain criterion performance. Another important finding that emerged from Experiment 6 concerns the gender by condition interaction in predicting retention. Males learned better from the LC condition than the other ones, while females performed better in the more stable conditions.

A pattern is beginning to emerge with regard to our reported interactions. In general, aptitude-treatment interaction (ATI) research has shown that certain learner characteristics are better suited to specific kinds of environments to achieve optimal outcome performance (see Shute, 1993a, 1993b; Shute, Glaser, & Raghavan, 1989; Tobias, 1989, 1994). For example, Shute (1993a) reported that individuals demonstrating greater exploratory behaviors perform better in more inductive or open kinds of learning environments (similar to the LC condition, and contrasting with more didactic ones) while the converse was found for less-exploratory individuals. A replication study (Shute & Gluck, in press) further refined exploratory behaviors in terms of on-line tool usage and reported gender effects related to tool use. That is, males tended to more spontaneously employ the on-line tools compared to females, and there was a main effect of tool use on learning outcome (i.e., more was better, overall). Finally, exploratory and independent kinds of behaviors have been linked to endogenous testosterone level, and males

have significantly more testosterone than females (e.g., Broverman, Klaiber, Kobayashi, & Vogel, 1968; Kimura, 1992; Newcombe, 1982). Testosterone affects brain functions in a manner similar to an adrenergic stimulant--exerting an influence on precisely those traits that are best suited to a learning environment offering more exploratory options and learner control.

Obviously, more research is needed in order to test all of these relationships. Within these two papers (Parts A and B), we've presented several intriguing condition by gender/testosterone interactions across a variety of domains. The implications of these findings are simple and straightforward--learning outcome and efficiency may be greatly enhanced by matching person traits to appropriate learning environment. Our preliminary studies have globally identified some treatment conditions that may serve as good points of departure for further, more refined research. Moreover, individual differences (showing up as gender effects) suggest different kinds of person-treatment pairings. Additional answers to the remaining, empirical questions arising from these six experiments will enable psychologists, educators, and instructional designers to subsequently adapt instruction even more precisely to the needs of individual learners.

References

Anderson, J. R. (1993). Rules of the mind. Hillsdale, NJ: Erlbaum.

Barlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press.

Broverman, D. M,. Klaiber, E. L., Kobayashi, Y., & Vogel, W. (1968). Roles of activation and inhibition in sex differences in cognitive abilities. Psychological Review, 75, 23-50.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), Cognition and instruction: Issues and agendas. Hillsdale, NJ: Erlbaum.

Collins, A. and Brown, J. S. (1988). The computer as a tool for learning through reflection. In H. Mandl and A. Lesgold (Eds.), Learning issues for intelligent tutoring systems (pp. 1-18), New York, NY: Springer-Verlag.

Corbett, A. T. & Anderson, J. R. (1989). Feedback timing and student control in the Lisp intelligent tutoring system. In D. Bierman, J. Brueker, & J. Sandberg (Eds.), Proceedings of the 4th International Conference on Artificial Intelligence and Education (pp. 64-72). Springfield, VA: IOS.

Gentner, D., & Gentner, D. (1983). Flowing waters and teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), Mental models (pp. 99-129). Hillsdale, NJ: Erlbaum.

Glynn, S. M. (1991). Explaining science concepts: A teaching-with-analogies model. In S. M. Glynn, R. H. Yeany, & B. K. Britton (Eds.), The psychology of learning science (pp. 219-240). Hillsdale, NJ: Erlbaum.

Goettl, B. P., & Shute, V. J. (in press). Systematic task decomposition of a desktop simulator using the backward transfer technique. To appear in the Journal of Experimental Psychology: Applied.

Kimura, D. (1992). Sex differences in the Brain. Scientific American, 119-125.

Kinzie, M. B., Sullivan, H. J., & Berdel, R. L. (1988). Learner control and achievement in science computer-assisted instruction. Journal of Educational Psychology, 80(3), 299-303.

Newcombe, N. (1982). Sex-related differences in spatial ability. In M. Potegal (Ed.), Spatial ability: Development and physiological foundations (pp. 223-250). New York, NY: Academic Press.

Piaget, J. (1954). The construction of reality in the child. New York: Ballentine Books.

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.

Shute, V. J. (1993a). A macroadaptive approach to tutoring. Journal of Artificial Intelligence and Education, 4(1), 61-93.

Shute, V. J. (1993b). A comparison of learning environments: All that glitters... In S. P. Lajoie & S. J. Derry (Eds.), Computers as cognitive tools (pp. 47-74). Hillsdale, NJ: Erlbaum.

Shute, V. J. (1995). SMART: Student Modeling Approach for Responsive Tutoring. User modeling and user-adapted interactions, 5, 1-44.

Shute, V. J. and Catrambone, R. (1996, July). Unified vs. tailored analogies: effects on conceptual knowledge acquisition. Proceedings of the ICLS 96 Conference, AACE, Washington DC.

Shute, V. J. & Gawlick, L. A. (1995). Practice effects on skill acquisition, learning outcome, and retention. Human Factors, 37(4), 781-803.

Shute, V. J., & Gawlick, L. A. (1996). The effects of instructional intervention, gender, testosterone, and stress on spatial learning. Technical Report, Air Force Material Command, Brooks Air Force Base, TX.

Shute, V. J. & Gawlick-Grendell, L. A. (1994). What does the computer contribute to learning? Computers and Education: An International Journal, 23(3), 177-186.

Shute, V. J., Glaser, R. & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences. New York, NY: W. H. Freeman, 279-326.

Shute, V. J., & Gluck, K. A. (1994). Stat Lady: Descriptive Statistics Module. [Unpublished computer program]. Brooks Air Force Base, TX: Armstrong Laboratory.

Shute, V. J. & Gluck, K. A. (in press). Patterns of Exploratory Behavior: Causes and Effects. To appear in The Journal of the Learning Sciences.

Sleeman, D., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. Cognitive Science, 13(4), 551-568.

Swanson, J. H. (1990, April). The effectiveness of tutorial strategies: An experimental evaluation. Paper presented the American Educational Research Association, Boston, MA.

Tobias, S. (1989). Another look at research on the adaptation of instruction to student characteristics. Educational Psychologist, 24(3), 213-227.

Tobias, S. (1994). Interest, prior knowledge, and learning. Review of Educational Research, 64(1), 37-54.