5-1992

Knowing Which Way Is Up: Sex Differences In Understanding Horizontality And Verticality

Gary A. Goodrich

Utah State University

KNOWING WHICH WAY IS UP: SEX DIFFERENCES IN UNDERSTANDING HORIZONTALITY AND VERTICALITY

by

Gary A. Goodrich

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Psychology

UTAH STATE UNIVERSITY
Logan, Utah

1992

ACKNOWLEDGEMENTS

I give my thanks to Sheryl, who has been warmly supporting during the first decade of our marriage, allowing me the necessary time to pursue my dream of getting a doctorate degree. I also thank Kevin, my son. He has filled my life with joy and given me a sense of magic and meaning that somehow escaped from me several years ago.

I express my gratitude to Dr. Ascione for his patience in reviewing and revising my thinking, writing, and research strategy. I'm most grateful to Dr. Ascione for teaching me that research can be fun. Although I had fairly good understanding of other aspects of the research process, this was a pleasant surprise.

I also thank Dave for being my best friend, philosophical cousin, research consultant, and fellow aficionado of silliness.

Gary A. Goodrich

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Knowing Which Way Is Up: Sex Differences In Understanding Horizontality And Verticality

by

Gary A. Goodrich, Doctor of Philosophy

Utah State University, 1992

Major Professor: Dr. Frank R. Ascione
Department: Psychology

In previous research men have been shown to obtain higher mean scores on tests of horizontality (H) and verticality (V) than do women. This study investigated the role of experiential factors in this gender difference. Undergraduate psychology students were randomly assigned to one of three treatment groups: training, enhanced training, and placebo. The pretest measure of verticality and horizontality (V/H), training, and posttest were administered via videotape. Major findings were: men obtained higher mean V/H scores than women; V and H scores correlated significantly; and training increased performance relative to a placebo, but enhanced training was not superior to standard training.

It was hypothesized that participation in athletics might eliminate the V/H sex difference. This was supported by initial analyses of the data. However, further analyses revealed that this may have been artifactual.

Errors on the V/H test were classified as undercorrections, overcorrections, and miscorrections. It was found that miscorrections corresponded to relatively low scores, inconsistent responding, and resistance to training. Scores on a generalization test substantiated all findings from the original V/H measure.

(162 pages)

INTRODUCTION

Piaget advanced a widely accepted and comprehensive theory of human cognitive development. Concrete operations, the third level in his stage theory, is characterized by a person's ability to reason and work with physical reality. Part of this ability entails a person's comprehension of three-dimensional, Euclidean space. (See Appendix A for a glossary of terms.) Piagetian theory holds that Euclidean concepts are mature when space is perceived in a three-dimensional coordinate axis system oriented by gravity. In order to meet this requirement a person must understand the concepts of vertical (any axis parallel to gravity's pull), and horizontal (any line or plane perpendicular to vertical). To test this understanding, Piaget and Inhelder (1956) devised tasks wherein children were asked to draw trees growing on hillsides and the surface of water in a tilted container. If the trees were shown in a vertical position and the water level was drawn horizontally, the child was said to have a mature concept of space.

According to Piaget, adults should exhibit facility with concrete operations. However, in studies examining adults' spatial concepts, surprisingly many people make errors indicative of preoperational thinking (Jamison & Signorella, 1980; Kalichman, 1986; Liben & Golbeck, 1984; Tegano, 1982; Thomas, Jamison, & Hummel, 1973). For example, many adults seem to believe that horizontal phenomena (like the surface of still water) and vertical phenomena (like plumb lines) are influenced by the orientation of surrounding structures. Such a misunderstanding could be very significant in tasks requiring concrete or formal operations, such as designing roads, buildings, and dams on sloped ground. Indeed, with much present attention and interest regarding equal participation by all people in math, science, and engineering, maturity of coordinate axis systems and Euclidean concepts becomes a highly important issue.

Women tend to make more errors on verticality and horizontality (V/H) tests than do men. Although there is much overlap, men often obtain higher mean V/H scores than do women. On a three-dimensional water surface measure used by Thomas and Jamison (1975), 13% of college men and 47% of college women made scorable errors. Kelly and Kelly (1977) used a paper and pencil measure of horizontal concepts. In their study 28% of men and 51% of women represented the surface of a liquid at least 10 degrees off of horizontal. Harris, Hanley, and Best (1978), on a water-level test, found a V/H sex difference among grade school children (85% of boys and 96% of girls making errors), and among college adults (35% of men and 57% of women made errors). Other researchers have acknowledged this consistent sex difference (Halpern, 1986; Liben, 1978; Liben & Golbeck, 1984; Peskin, 1980).

There have been many attempts to account for observed sex differences in V/H measures. According to Kalichman (1988) these include reference to biological characteristics (brain structure, timing of puberty, hormonal influences), cognitive-behavioral variables (Piagetian stages, interaction of cognitive stage and age), and experiential factors (sex roles, training, performance expectancies). Experiential factors, in particular, have received much attention in recent research (Kalichman, 1986; Newcombe, Bandura, & Taylor, 1983; Petersen, 1983). Two questions are asked regarding the role of experience in V/H sex differences. First, what types of experience might influence the development of this sex difference? Second, what forms of experience might remediate poor V/H test performance?

One type of experience that might account for some of the V/H sex differences is participation in sports (Petersen, 1981). The effect of athletic experience on V/H understanding is an especially promising research area since currently males are more likely to participate in sports than are females and because physical activity correlates with performance on the Rod and Frame Test (Svinicki, Bundgaard, Schwensohn, & Westgor, 1974) and other spatial tasks (Olson, Eliot, & Hardy, 1988; Newcombe et al., 1983). Should athletics play a meaningful role in the development of mature coordinate reference systems, it might account for a substantial portion of the known sex difference in V/H test scores. Unfortunately, the relation between athletic experience and V/H sex differences has never been directly examined.

Even if it is found that athletic experience correlates with V/H sophistication, we would still be left with the question of how to correct or improve faulty understanding of vertical and horizontal invariance. In an effort to address the trainability of verticality/ horizontality, there have been several formal and informal attempts to teach people about water levels and plumb lines. Most of these efforts have met with failure or nongeneralizable success (Barsky & Lachman, 1986; Beilin, Kagan, & Rabinowitz, 1966; G. N. Kelly & J. T. Kelly, 1978; Liben, 1978; Thomas et al., 1973).

Programs designed to train V/H abilities have included several discrete approaches. Guided physical contact with half-full containers (Thomas et al., 1973), plumb lines (Meehan & Overton, 1986), physics lectures and heuristics about gravity (Liben & Golbeck, 1984), and explicit instructions (Beilin et al., 1966; Liben & Golbeck, 1984) have all been used. But in all training efforts there have been many subjects who failed to reach levels of performance predicted by Piagetian theory.

An exciting possibility is that many subjects who fail certain Euclidean tests might actually have intact coordinate reference systems (Blades & Spencer, 1989). That is, people may demonstrate Euclidean accuracy in certain testing situations, but fail when presented with traditional water-level and plumb line tasks (Liben & Golbeck, 1980, 1986; Meehan, 1984; McGillicuddy-De Lisi, De Lisi, & Youniss, 1978; Peskin, 1980). Although there have been repeated indications that adults may have accurate but hidden coordinate reference systems, no training programs have attempted to help adults make the conceptual connection between their existing knowledge and the V/H tasks used in testing.

Despite evidence that athletic experience might correlate with Euclidean concept maturity, researchers have not directly assessed the relation between sports participation and V/H understanding. Despite several promising possibilities for effectively training V/H comprehension, little research has been conducted to assess the possibility that some people may actually have accurate, but untapped, coordinate reference systems. The present is a report of research into the relation between V/H comprehension and athletic experience and into the effects of a V/H training program designed to access and harness hidden, but accurate, Euclidean comprehension.

Objectives

The overall goal of this research was to increase understanding of how experience relates to sex differences in V/H concepts. The research was guided by the following specific objectives:

1. To determine whether V/H training success might be enhanced by attempts to activate and utilize people's accurate but often neglected knowledge of V/H phenomena (e.g., standing on a hill or drinking from a glass).

2. To determine whether there is a V/H sex difference among adults who have extensive experience participating in athletics.

In order to attain these objectives it was necessary to meet several goals. These included:

1. Measuring verticality and horizontality simultaneously and accurately.

2. Identifying men and women who have extensive athletic experience.

3. Identifying methods of training V/H concepts in previous research.

4. Identifying ways to activate Euclidean concepts in people who fail standard V/H tests.

The following is a report of the methods, rationale, and results of the research effort that was guided by these objectives.

REVIEW OF RESEARCH

In this literature review, the relation of horizontality and verticality will be discussed first. Then literature directly relevant to V/H sex differences will be examined. Next, the focus will turn to the possibility that people may have accurate understanding of coordinate reference systems, but may fail to apply their knowledge when appropriate. Finally, correlates of V/H scores and possible mechanisms for a V/H sex difference will be discussed.

The Correlation Between Verticality and Horizontality

Verticality and horizontality are correlated theoretically, conceptually, and empirically.

Relation of Verticality and Horizontality in Piagetian Theory

Piaget and Inhelder (1948/1956) examined the coordinate reference aspect of children's spatial concepts by assessing their understanding of vertical phenomena (like trees on hills and plumb lines) and horizontal phenomena (like water levels). Their discussions of cognitive stages and patterns of development were applied to both verticality and horizontality. Piaget and Inhelder argued that the two concepts developed simultaneously and together lead to mature Euclidean concepts. In Piagetian theory, then, verticality and horizontality are related.

Conceptual Relation of Verticality and Horizontality

Euclidean space is hypothetical, three-dimensional space. In the Euclidean scheme there are three mutually perpendicular axes which extend ad infinitum. One axis is vertical, another is horizontal, and a third represents depth. The vertical axis is fixed in accordance with gravity. The other two dimensions are constrained to fall within a horizontal plane.

There are situations in which Euclidean space is neither accurate nor useful. For example, on a global scale it is apparent that not all plumb lines are parallel and that the horizon is not planar. On the smaller scale more typical of daily human experience, however, Euclidean space is a practical and economical means of conceptualizing the physical world.

In the Euclidean space of human experience horizontal and vertical phenomena are invariantly perpendicular to each other. Both are determined by the pull of gravity. Neither is influenced by the slope of nearby structures. Therefore, verticality and horizontality are related to each other in that they are part of Euclidean space, derive from the same physical basis (gravity), are mutually perpendicular, and are unchanged by the slope of nearby structures.

Empirical Relation of Verticality and Horizontality

Adults' verticality and horizontality scores often correlate at a statistically significant level. Mackay, Brazendale, and Wilson (1972) found significant correlations of .39 for men and .58 for women on vertical and horizontal scores. Liben (1978) gave a water level test, tree-on-hill test, and plumb line test to 12th grade students. The correlation between sum of tree and plumb lines tests (verticality) with water level was .77 for 30 boys ($p < .05$) but only .22 for 28 girls ($p > .10$).

Liben and Golbeck (1984) had college students complete standard V/H tasks (plumb line and water level) and nonphysical V/H tasks. For these latter tasks subjects were asked to draw lines "straight across" or "straight up and down" within oblique rectangles. V/H correlations were statistically significant for men and women on all variants of the tasks. In later research Golbeck (1986) again found that college students' scores on water level and plumb line tasks were significantly correlated ($r = .81$ for 32 men; $r = .45$ for 32 women).

Ohuche (1984) gave a standard water level test, a novel horizontality task, and a verticality task to 192 Igbo (Nigerian) public school students, ages 8 to 19 years, selected through a stratified random sampling procedure. The verticality task required subjects to draw lines representing posts which were to be set "nice and straight" on hills. The novel horizontality task involved a tray balanced on a woman's head. Six drawings of a profile of a woman's head, obliquely positioned, were presented. Subjects were asked to draw lines indicating a tray balanced on her head.

Two peculiarities limit the value of the balanced tray measure. First, human head profiles have no consistent straight line referents to indicate orientation, so obliqueness is not inherently meaningful. Second, a flat tray need not be horizontal to remain balanced on a rounded object.[1]

Subjects received brief training specific to the three measures prior to collection of the dependent variable data. Intertask correlations were computed separately for males and females. All were statistically significant, ranging from .45 to .57 ($p < .001$ in all cases).

The Rod and Frame Test (RFT) is often used as a measure of field articulation, or cognitive style. The dependent variable in the RFT is the subject's accuracy in orienting a rod vertically within an oblique frame. Barsky and Lachman (1986) administered the RFT, a standard plumb line task, and a water level task to 68 undergraduate college women. All intercorrelations were statistically significant, ranging from .30 to .42. Willemsen, Buchholz,

---

[1] The present author tested this informally, using a tacky rubber ball and a book. The book remained balanced even when deviating 35 degrees from horizontal!

Budrow, and Geannacopoulos (1973) found a Spearman rho correlation of .37 between water level scores and RFT for 30 undergraduate college women.

There is ample evidence to conclude that horizontality and verticality are empirically related. This is true whether the V/H scores are derived from standard Piagetian tasks such as water levels and plumb lines, less common tasks such as tray balancing and pole setting, or from conceptually related measures like the RFT.

Gender as a Variable in V/H Tests

There is much published evidence of a male advantage on V/H measures, but only one study will be described here in detail. It is the classic, serendipitous study that originally stirred interest in a V/H sex difference.

To demonstrate the notion of horizontality to college psychology students, Rebelsky (1964) gave a water-level test to 59 men and 69 women (ages from 20 to 26 years). They were shown five pictures of rectangular glasses, one upright and the others tilted to various angles. Subjects were asked to draw lines showing what the water would look like if the glasses were half full. Lines drawn within five degrees of horizontal were counted as accurate.

Men obtained significantly higher scores on two of the figures, but there were no statistically significant differences on the other two tilted glasses. After debriefing, many students were still perplexed by the claim that still water is always horizontal. Some students obtained glasses to experiment with the phenomenon. Even after verifying the principle, some students argued that it still "looked" tilted.

Rebelsky marvelled that adults, with 20 years of experience drinking from tilted containers, could be so thoroughly ignorant of the principle on which the phenomenon rests. This was one of the first studies to note a sex difference in coordinate reference systems and to provide evidence that exposure to a physical law does not necessarily result in mature Euclidean concepts.

Rebelsky's report foreshadowed a common reaction among V/H researchers: dismay that so many people could misunderstand such stable, simple, common phenomena. For example, Liben and Golbeck (1984) wrote, "Surprisingly poor performance on Piagetian horizontality and verticality tasks has been found among late adolescent and young adult subjects" (p. 596). G. N. Kelly and J. T. Kelly (1978) observed:

> The reader is left to speculate how failure to understand a concept so basic might affect understanding of more sophisticated concepts. The reader must also wonder, as do the authors, how many other concepts essential to understanding of the physical world are not understood by a large proportion of females. (p. 31)

There have been many studies since Rebelsky's initial report which have examined the possible V/H sex difference. Most of these studies strengthen her tentative finding that, as a group, men obtain higher V/H scores than women.

Some of the researchers who have found a significant sex difference in adults' scores on V/H tasks are Golbeck (1986), Harris et al. (1978), Jamison and Signorella (1980), Kalichman (1986, 1987, 1989) G. N. Kelly and J. T. Kelly (1978), J. T. Kelly and G. N. Kelly (1977), Krasnoff, Walker, and Howard (1989), Liben (1991), Liben and Golbeck (1984, 1986), McGillicuddy-De Lisi, De Lisi, and Youniss (1978), Meehan and Overton (1986), Morris (1971), Signorella, Jamison, and Krupa (1989), Thomas and Jamison (1975), Walker and Krasnoff (1978), Willemsen and Reynolds (1973), and Wittig and Allen (1984). Although the causes, correlates, and implications of the sex difference are being hotly debated, the preponderance of evidence regarding male advantage in V/H tasks has led to its general acceptance as an established fact.

Task Variables and Sex Differences

Many researchers have studied how different aspects of V/H tasks influence performance. Shape, orientation, and content of the test stimuli have all been examined for their influence on V/H scores. In general these variables have been shown to affect performance of males and females, but not contribute significantly to the observed sex difference.

Shape of Test Stimuli and V/H Scores

In a study on how container shape affects horizontality scores, Willemsen and Reynolds (1973) presented 30 college students with water-level tasks involving three shapes of containers: petri dish, cylindrical "shampoo" bottle, and florence flask.

The test apparatus was a glass container cut in half, mounted in front of a vertical surface. The container could be rotated 360 degrees. Visible through the glass was a disk painted to represent a liquid surface. This disk could be rotated by the subject to any angle. At the side of the test container was a model--an actual container half filled with colored water. Like the test container, the model was mounted in front of a vertical surface and could be rotated. The model could be covered by a form-fitting slip.

After being introduced to the apparatus, subjects watched as the model (petri dish, florence flask, or shampoo bottle) was covered and rotated. The test bottle was rotated to the same orientation. Subjects were asked to set the bicolored test disk to show how the liquid in the model would look. The outcome measure was the deviation from horizontal (in degrees) at which the subjects set the test disk.

Men and women made fewer errors on trials involving the petri dish than on trials with the shampoo bottle or flask. Men made fewer errors than women on flask and bottle trials. For both men and women, oblique presentations of the flask and bottle evoked more errors than did orthogonal presentations. Willemsen and Reynolds (1973) concluded, "The sex differences in adults' performance on the water-line task result from the greater tendency of females to be influenced by the straight-line characteristics of the containers" (p. 309).

Liben (1978) gave two water level tasks to 33 male and 33 female 12th grade students. One task used a rounded water container. In the other task the container was rectangular. Black line drawings of each shape were first shown upright with a correctly rendered liquid line. Then each shape was shown in five oblique orientations. Subjects were asked to draw in the water line. Any line within 10 degrees of horizontal was scored as accurate.

Verticality was measured by a plumb line task and a tree task. For the plumb line, subjects were shown an outline drawing of a trailer with a light bulb dangling from a cord in the ceiling. They were also asked to draw the bulb and wire in pictures of the trailer set on hills sloping 30, 45, and 60 degrees.

Data analyses showed no significant effect for container shape on horizontality scores. Verticality was higher when measured by renderings of trees than when measured by plumb lines. Water levels were more accurate for orthogonal test stimuli than for oblique items. But there was no interaction between sex and task for either verticality or horizontality.

Other researchers who studied the effect of container shape have found no interaction with sex. For example, Thomas and Jamison (1975) found that both boys and girls in grade school made fewer errors when using a rounded flask than when using a rectangular bottle. Murphy-Berman, Witters, and Harding (1986) found a similar effect for hearing-impaired students ranging in age from 9 to 12 years: A rounded flask induced fewer errors than did a rectangular bottle. The effect of container shape, however, was not mediated by subject gender.

Orientation of Test Stimuli

Container tilt influences scores and may interact significantly with sex. Wittig and Allen (1984) compared scores on three horizontality measures: multiple choice format, drawing a line in a container, and manipulating an apparatus. A four-way mixed design analysis of variance revealed significant main effects for sex (men had higher means), test format (apparatus manipulation yielded higher scores than line drawing or multiple choice), and container orientation (orthogonal stimuli were easier than oblique stimuli). Interactions were significant for orientation with sex and for orientation with task. A principal axis factor analysis showed that all three instruments tapped a single factor.

Ceiling effects were pronounced on orthogonally oriented test stimuli, with almost no variance as nearly all subjects obtained perfect scores. This ceiling effect was so strong that the authors recommended future research into V/H sex differences only include oblique test stimuli. They wrote, "Since sex and task differences in mean performance appear in angled trials only, researchers may wish to delete horizontal and vertical trials" (Witting & Allen, 1984, p. 311).

Other researchers have found that oblique test stimuli produce lower V/H scores than do orthogonal stimuli. Willemsen and Reynolds (1973), discussed above, found that test stimuli which are either vertical or horizontal lead college students to produce very few errors. Similar results were found by Smedslund (1963) for children ranging in age from five to seven years and by Beilin et al. (1966) for children six to eight years of age.

Content of Test Stimuli

Peskin (1980) observed that much research into Piaget's formal and concrete operations stage has shown a sex difference, often favoring males. She noted that many of the tasks used by the researchers were scientific in nature and may have been more interesting to males than to females. Peskin devised several new tasks with content judged to be more feminine (dealing with cosmetics, for example) and substituted these for some of the standard Piagetian tasks.

Peskin's subjects were 101 South African girls (ages 14 to 16) who had either studied science in school or who had nonscience education backgrounds. For science students, the content of the tasks (standard, scientific tasks or novel, feminine tasks) did not influence their scores. For nonscience students, however, the content significantly influenced the cognitive stage they obtained on the tests. That is, nonscience students obtained higher rated cognitive levels on kitchen and cosmetic tasks than on pendulum or chemistry tasks. Peskin argued that experiential factors and task content must be taken into account before sex differences can be adequately explained.

Ohuche (1984; described previously) gave test-specific training prior to testing 192 Nigerian students. She found that males obtained higher means on two horizontality measures, but not on a posts-on-hill verticality task. As with Peskin's research, this raises the interesting possibility that task content may be a factor in the common finding of a significant V/H sex difference. Of course, the fact that the dependent variable was collected after training limits the inferential value of the results.

De Lisi (1983) observed that two main hypotheses have been used to account for V/H sex differences: competence deficit (people fail V/H tests because they don't have the concepts) and deficits in competence utilization (they know, but fail to apply their knowledge in certain settings). To help determine which hypothesis is more valid, 60 children (10 boys and 10 girls in first, third, and fifth grades) were tested on the RFT, water level test, and a crossbar horizontality test. The crossbar apparatus was a metal bar which freely pivoted on the end of a rod which could be set to any angle. The crossbar was weighted to remain horizontal regardless of the position of the rod. Subjects were allowed to manipulate both water level and crossbar tasks before testing. For testing, the apparatus was hidden and set to a specific orientation. Subjects were given an outline of the bottle or rod set to that orientation. Their task was to draw a line representing the crossbar or water level. Corrective feedback was given after each of six introductory trials. For testing no feedback was provided.

De Lisi argued that the water bottle task has a stronger field effect than the crossbar task. If the competence-deficit hypothesis is more accurate, then two patterns should emerge in the data. First, field-independent subjects should outperform field-dependent subjects on both horizontality tasks. Second, correlations between horizontality measures should be equal for field-dependent and field-independent subjects.

Analyses showed that field-independent subjects outperformed field-dependent subjects on both tasks, although the difference was larger for the water level task. Overall crossbar scores were higher than water level scores. Effect sizes computed from means and deviations showed that boys had higher means on both the water level task (*d*=.62, .26, and .46 for first, third, and fifth grades, respectively) and crossbar task (*d*=1.13, .56, and .50). Despite this, the results were seen as supporting the competence-utilization hypothesis.

Confounding factors make interpretation of this research difficult. First, cognitive style was assessed by the RFT, a measure of verticality within an oblique frame. In its reliance on verticality the RFT is conceptually entangled with the dependent variables.

A second confound arises from the effort to reduce oblique distractors. Without oblique stimuli one cannot adequately assess the upper limits of V/H understanding. In the final analysis, then, this study tells us that people who find it difficult to indicate vertical in the presence of oblique stimuli also find it difficult to indicate horizontal in the presence of oblique stimuli.

Golbeck (1986) hypothesized that performance variables such as physics knowledge, not competence in V/H knowledge, account for the usual V/H sex difference. To test this, Golbeck administered paper and pencil tasks to 32 men and 32 women undergraduate psychology students. The tests were either physical (water level or plumb line) or nonphysical ("draw a line straight across" or "draw a line straight up and down"). All lines were to be placed in similarly sized, shaped, and oriented rectangles. In the physical version the shapes were said to represent a glass or a van. For the nonphysical tests, subjects were shown a correctly solved problem involving a stimulus tilted 45 degrees. No solved oblique examples were available for the physical versions.

Analyses of variance showed significant main effects for sex (men had higher mean scores) and task (nonphysical tasks produced higher scores), but no significant interaction. Data were reanalyzed, this time comparing the number of men and women who passed the tests (scores of five or six out of six items). Chi square analyses showed that a higher proportion of men than women passed the physical V/H problems, but that there was no difference in proportions passing the nonphysical tests. Golbeck concluded that the data support the hypothesis that physical knowledge, not V/H competence per se, underlies the sex difference.

It should be noted that the data may be interpreted as supporting a converse conclusion. Golbeck had predicted that a significant sex by task interaction would be present if the underlying cause of the V/H sex discrepancy was knowledge of physical phenomena. There was no interaction of sex with task for either horizontality or verticality. In addition, the proportion of women who failed the nonphysical V/H tasks (getting scores of two or less) was at least double the proportion of men. Finally, the procedures were different for the two tests. Nonphysical tests began with an example of a correctly completed oblique trial. No such guide was given prior to physical tests. In summary, then, it is unwarranted to rule out differential competence as a cause of the V/H sex difference.

Summary of Task Effects

Although many studies have found significant task effects, these effects did not interact significantly with sex. Thomas and Jamison (1981) interpreted this to suggest "that the sex differences on the water-level task cannot be explained on the basis of task characteristics alone" (p. 275). The research reviewed here supports their conclusion. Despite strong evidence that task variables may influence V/H scores in predictable ways, there is very little evidence of significant or replicable interactions between task variables and sex.

Many of the discrepancies between the studies discussed above can be explained by the concepts of task difficulty (as mediated by obliqueness, for example) and different V/H population curves for men and women. If a normal distribution of V/H skill is assumed, and if men demonstrate higher V/H skills than women, then studies which incorporate difficult V/H tasks should consistently find sex differences. If the V/H measures are very easy, producing limited variance and strong ceiling effects, then significant differences between sexes will be obscured. Researchers who found no sex difference in adults on certain V/H tasks (crossbar task in De Lisi, 1983; and nonphysical V/H tasks in Golbeck, 1986) utilized measures yielding scores much higher than typically found in V/H research. That is, they were very easy tasks, with nearly all subjects obtaining perfect scores. A notable exception is Ohuche (1984), whose posts-on-hill task produced scores equivalent to a water level task, but without a significant sex difference. It is at least plausible, then, that task difficulty and differential V/H skill curves for men and women may account for otherwise conflicting research findings. There is some empirical evidence that addresses the issue directly.

Liben (1991) expressed puzzlement over the results of crossbar apparatus research. So she developed systematic variations on the task used in De Lisi (1983) and McGillicuddy-De Lisi et al. (1978) to determine why their crossbar tasks produced no significant sex differences and infrequent errors by adults. Liben hypothesized that the original crossbar task had an insufficiently oblique frame of reference.

Fifty men and 54 women introductory psychology students completed the original (disembedded) crossbar task, a second task in which the crossbar apparatus was placed within a rotatable frame (embedded crossbar), the Embedded Figures Test (EFT), and a water level test. If poor performance was due to knowledge of the physics of liquid surfaces, then the water level task would be expected to yield lower scores than the crossbar tasks. If poor performance resulted from oblique context, then the water level task and the embedded crossbar task should be equally difficult relative to the disembedded crossbar. If field dependence were a key factor, then it should interact with task and sex effects.

Responses within five degrees were scored as correct on all horizontality tasks, with possible scores ranging from zero to six. Subjects were classified as field dependent or independent based on their position in a median split of EFT scores. Field-independent subjects scored nonsignificantly higher than field-dependent subjects on horizontality tasks. Field dependence did not interact with other factors. Embedded crossbar and water level scores were not significantly different from each other, but were lower than scores on the disembedded (original) task. The results supported the hypothesis that oblique context, not specific knowledge difference about the physical properties of water, accounts for the sex difference.

The crossbar pivoted on its visual center point and remained horizontal only because of hidden weights. It was hypothesized that this might have perplexed people who were familiar with the relevant physics. A second crossbar was built, this time with the pivot hole being mounted above the actual and apparent center of gravity. Embedded and disembedded crossbar tasks and water level task were administered to 100 men and 100 women introductory psychology students. After completion of the tasks they were asked to write how they had solved the problem.

The main effects of task (embedded or not) and sex were statistically significant. The interaction was not. Effect sizes were .90 on water level (favoring males), .87 on embedded crossbar (again favoring males), and .00 on the disembedded crossbar. Analysis of subjects' written explanations suggested that thinking in terms of the embedding frame lowered scores, while thinking in terms of external factors (gravity, the paper, etc.) led to higher scores. Taken together the two studies support strong inference that the relatively high scores on the original crossbar tasks resulted from the use of an insufficiently oblique embedding frame. Put more simply, crossbar studies only show that adults succeed on horizontality tasks when there is little oblique distraction. The studies do not point to situationally specific knowledge of horizontality.

The Role of Experience in V/H Knowledge

<u>Training</u>

Thomas et al. (1973) attempted to teach college women that still water is invariantly horizontal. Subjects were placed in front of an apparatus with two bottles: the "model," half-filled with red water, and the test figure, half a bottle which had been bisected vertically. The half-bottle was mounted in front of a red and white disk which could be rotated. For testing, the model was covered and the disk was rotated at least 45 degrees. Orientations of test trials corresponded to oblique clock positions. Subjects adjusted the disk to predict the fluid level in the model. Then the cover was removed and the water level and prediction were compared. Subjects whose responses were five or more degrees from horizontal on two or more trials were termed "naive."

Thirty of 47 women college students were found to be "naive" about the task. They were trained by repeating the task as many as 48 times. Only seven of the subjects met the criterion of successful learning (to pass 10 consecutive trials during 48 training trials.)

A second sample of 33 horizontality-naive women received similar training, only this time the model wasn't covered. Following 24 training trials, eight test trials were conducted. Again, few subjects exhibited successful learning. The authors concluded that ". . . subjects who perform inaccurately do so because they lack conceptual understanding that still water remains horizontal" (p. 174).

This study substantiated Rebelsky's (1964) observation that the principle of horizontality is not readily learned by adults who don't already understand the concept.

G. N. Kelly and J. T. Kelly (1978) conducted several V/H training studies. In the first of four studies, 314 university education students were shown a bottle, half-filled with colored water on which floated a small raft with a vertical mast. Subjects were then shown 12 drawings of the bottle rotated to the 12 clock positions and were asked to imagine it being slowly rotated. On the test sheet was a grid of faint vertical and horizontal lines. The subjects' task was to draw the water surface. More men (72%) than women (49%) passed the test (making no error greater than five degrees and a sum of errors less than 30 degrees).

Similar testing was applied to 158 boys and 120 girls ranging in age from 9 to 12 years (J. T. Kelly and G. N. Kelly, 1978). An analysis of variance showed a significant effect of gender for the 10-, 11-, and 12-year-olds, but not 9-year-olds. Kelly and Kelly tried to improve performance by providing fourth graders with a "science learning center," a collection of materials and instructional guides designed to provide interesting activities regarding horizontality. The center was placed in the back of a classroom for one week. Then posttesting was conducted. A Wilcoxin signed-ranks test for matched pairs showed that the class with the activity center made significant gains, while another fourth grade class, which had no such center, did not.

Finally it was decided to try teaching 29 college women the horizontality concept. They were given the original water-level test and were asked to state the relevant principle. Then subjects were tested using a novel, three-dimensional test. Several glasses had been tilted while colored gelatin set inside. Subjects were asked to place the glasses onto a lump of clay and adjust them to proper orientation. After completing the task, subjects were asked again to state a principle that guided their performance. Fourteen subjects were unable to state an accurate rule and were retained for training. They were given guided experience with several three-dimensional models and encouraged to state rules or principles that might apply. Compared to a control group, the treatment group showed substantial learning. However, eight of the 14 subjects did not learn the principle.

Liben and Golbeck (1984), described earlier, postulated that knowledge of physical phenomena, not underlying Euclidean competence, might account for sex differences on verticality and horizontality tasks. To eliminate the effect of differential physics knowledge, Liben and Golbeck planned an intervention wherein subjects were given the relevant physics information prior to the tests. The results, they claimed, would thus bypass physics awareness and focus on subjects' spatial concepts.

In the first of two experiments, 80 college men (mean age=18.8 years) and 80 college women (mean age=19.7 years) were randomly assigned to one of four conditions and given plumb line and water-level tests. Subjects in the first group were told a rule about how to correctly complete the task. A second group was shown an example of the task correctly completed. A third group was given the rule and an example. The control group received no prompting. Dependent variables were scores on plumb line and water-level tasks. Only the main effect of sex was statistically significant. Presentation of the rule, the example, and their interaction were not significant. There was no support for the hypothesis that sex differences in V/H concepts results from knowledge, not competence deficits.

In a second experiment Liben and Golbeck (1984) randomly assigned 80 college men and 80 college women to treatment and control conditions. This time treatment subjects were given more explicit rules for plumb lines and water levels. For example, regarding water levels they were told:

> To answer these problems correctly, it is important that you know [that] water remains horizontal or level, regardless of the position of the container. Remember, even though the glass is tipped, the water line will be straight across, or horizontal. (p. 601)

Subjects were quizzed on the rules and then tested. Scoring was the same as in experiment one.

An analysis of variance yielded significant effects only for sex. Because the main effect of task was nearly significant, the data were reanalyzed with Kolmogorov-Smirnov nonparametric tests. These analyses showed that in the control group a significantly higher proportion of men than women passed (scored 5 or 6) the tests. However, when the relevant information was first imparted to the subjects, the sex difference was statistically nonsignificant. The authors concluded that the data support their original hypothesis that physics knowledge, not underlying competence, explains sex differences on these tasks.

These two experiments provide valuable evidence regarding Euclidean concepts, but suffer procedural confounds. When the first experiment did not support the hypothesis, a second experiment with a more powerful treatment was conducted. When the hypothesis was again not supported, the data were reanalyzed with less powerful statistics, bringing the sex difference for the treatment group below the significance threshold. The significant/not significant dichotomy was then interpreted as supporting the initial hypothesis.

It might be argued that the treatment explicitly described not only plumb lines and water surfaces, but also trained to the outcome measure. This criticism, however, misses the authors' point that the instruction could not have been effective unless subjects accurately understood "up and down" and "straight across." This understanding, they suggested, is the essence of a Euclidean referent system.

Most efforts to train V/H concepts have met with small, immediate gains. However, in most studies there are some subjects who simply do not grasp the concept, no matter how energetically, creatively, or directly it is taught (G. N. Kelly & J. T. Kelly, 1978; Liben, 1978; Liben & Golbeck, 1984; Thomas et al., 1973) Additionally, the gains often fail to generalize to related measures. For example, Barsky and Lachman (1986) produced significant gains in water level scores for 49 college women who failed an initial water level test. But when the subjects were retested on a water level task using a different bottle shape, the gains disappeared. A similar generalization failure was found for 152 second grade students who received perceptual and conceptual horizontality training (Beilin et al., 1966). Subjects showed gains on the specific water level task which was the focus of training, but failed to demonstrate generalization of learning when tested with a bottle of a different shape.

Athletics and Physical Activity

Svinicki et al. (1974) gave a one-trial version of the RFT to five physically active men, five inactive men, five active women, and five inactive women. Subjects were classified as physically active if they reported spending five or more hours weekly in "strenuous physical activity," classified themselves as physically active, and reported a history of strenuous physical activity. Criteria for inactivity were spending less than one hour per week in strenuous physical activity, self-classification as inactive, and a history of inactivity. No details were given about which specific activities were involved.

An analysis of variance for the dependent variable (degrees deviation from vertical on the Rod-and-Frame test) showed no significant sex difference, but did yield a significant effect for physical activity. Comparisons of mean scores showed that physically active subjects had higher scores than physically inactive subjects. Regarding sex differences in field-dependence Svinicki et al. (1974) concluded:

> [Sex differences] might be related to the relative likelihood of males and females engaging in physical activity. Females who are more physically active should be more field-independent and the sex difference might disappear. (p. 1238)

Newcombe et al. (1983) listed adolescent and adult activities that involve spatial skills and sex-typing. Activities such as hunting, frisbee, air hockey, juggling, drawing, navigating a car, ballet, quilting, and typing were categorized as masculine, feminine, or neutral. The final list of 81 sex-related spatial activities was selected on the basis of interrater reliability in categorizing the activities as masculine, feminine, or neutral. There were 40 masculine items, 21 feminine items, and 20 neutral activities. The Spatial Activities Questionnaire (SAQ) was administered, in checklist form, to 22 men and 23 women college students for whom Differential

Aptitude Test (DAT) scores were available. Scores on the SAQ correlated .33 with all subjects' scores on the Spatial Relations portion of the DAT (for women r=.40; for men r=.18). It should be noted that 26 of the 40 "masculine" activities and 3 of the 21 "feminine" activities might be classified as sports. Thus it appears that participation in sports seems to correlate with high scores on at least one measure of spatial abilities.

In a partial replication of the above study, Olson et al. (1988) investigated the correlation between spatial abilities and physical activities for 53 college women and 45 college men. Assessments included the Spatial Dimensionality Test (an experimental battery of tasks like hidden figures, mental rotations, and paper folding), the Spatial Antecedents Questionnaire (measuring spatial activities such as sports and hobbies), the Academic Courses Scale (to measure experience with spatially-relevant academic courses), the Self-assessment Scale (subjects rate their skills on spatial tasks relative to their peers), the Environment Mapping Scale (knowledge of campus landmarks, judgment of distance between landmarks, and awareness of direction), and the Revised-Individual Questionnaire (assessing subjects' preference for visual or verbal processing style).

Previous course work and self-assessed capacity accounted for most of the variance in spatial performance. For the total sample it was found that participation in sports such as football, skiing, soccer, and bicycling correlated positively with spatial scores. Surprisingly, for women and for the entire sample there was a significant negative correlation between participation in ballet and choreographing dance and SDT scores. It should be noted that women rated themselves as highly as did males on the self-assessment of spatial abilities, in contrast to some previous research (Lunneborg & Lunneborg, 1984).

This research suggests that participation in spatial activities may correlate with scores on spatial tests. In that context, athletics may be correlated with spatial skills.

Signorella et al. (1989) attempted to shorten the 81-item SAQ (described above; Newcombe, et al., 1983) to 10 masculine items (e.g., baseball, football, target shooting), 10 neutral items (e.g., tennis, diving, bowling), and 10 feminine items (e.g., figure skating, embroidery, quilting). This short version of the SAQ included a seven-point rating of frequency of participation in each activity. Alpha coefficient reliabilities for the masculine and feminine subscales were .79 and .77, respectively (Signorella, Krupa, Jamison, & Lyons, 1986). No validity data were available. The shortened SAQ, Card Rotation Test (CRT), Bem Sex Role Inventory (BSRI), and a paper-and-pencil water level task were given to 146 women and 198 men in a general psychology class. On the water-level test men outscored women ($\underline{d}$=+.53). In a causal modeling correlational analysis there was a significant effect of SAQ for men and women, while the BSRI interacted with water-level for women, but not for men. The authors drew the following conclusion:

> The effect of BSRI M [masculine score] was not mediated by activity participation. This suggests that, although individuals with higher BSRI M scores are more active, it is masculine activity participation, not activity in general, that relates to spatial performance. (p. 95)

They suggested that future research focus on participation in masculine-stereotyped activities, particularly during adolescence.

Petersen (1981) reviewed sex differences in spatial performance. She noted that spatial performance, field independence, and horizontality are related concepts (a position later recanted; Petersen, 1983). The review produced three main factors that might account for sex differences in spatial abilities: parental socialization, sex role socialization, and biological factors. Petersen (1981) then conducted original research to test whether these factors might significantly account for sex differences in spatial abilities.

Extreme groups sampling yielded a matrix of six cells, each with 25 suburban high school seniors. The cells were based on sex and Guilford-Zimmerman Clock Test scores (high, middle, or low). The BSRI, Parent Behavior Form, self-report of timing of puberty, and a dichotic listening task were administered to all subjects.

There was no evidence that boys and girls differed in their involvement with parents, although high parental involvement coincided with high spatial scores for both boys and girls. No within groups sex role differences were found, failing to support a sex role/spatial abilities correlation. There was no effect for dichotic listening. Subjects with low spatial abilities began puberty at an earlier age than middle and high spatial abilities subjects. Athletic boys scored higher on the spatial measures than nonathletic boys. This relation did not hold for girls. The biological factor of puberty onset and athletic participation correlated with spatial abilities. It is difficult to interpret the interesting findings of athletics within the context of this study since athleticism can not be said to result from only one of the posited factors: parental socialization, sex role socialization, or biological effects.

As is evident in the research described above, there are several reasons to anticipate that sports participation might correlate with V/H test scores. First, there are many sports which require superior V/H comprehension. Any golfer who fails to account for the slope of a green would be a pitiable putter. Successful participation in balance-intensive sports such as gymnastics, skiing, and figure skating rests on precise awareness of the vertical pull of gravity. A gymnast on a balance beam would fall repeatedly if she tried to stand perpendicular to the slope of nearby bleachers. Likewise, a skier who anticipates that the pull of gravity is influenced by the angle of the hill would be incapacitated.

The correlation between athletics and V/H skills has not been empirically established. Indirect evidence comes from research dealing with spatial abilities other than horizontality or verticality. These data indicate that spatial abilities correlate positively with participation in sports and related activities. The only evidence that directly addresses the issue of V/H abilities and physical activity (Svinicki et al., 1974) suggests that a significant relation exists. Although there is reason to anticipate a correlation between sports and V/H abilities, little research into the matter has been conducted.

Schooling

Kalichman (1989) argued that V/H research conducted with college students ought to take into account the subjects' academic major. Kalichman (1986, 1987) administered paper-and-pencil water level tests to undergraduate students in five college majors. There were 25 women and 25 men in each of the majors (N=250). Four straight-sided bottles and four round bottles were pictured at 30, 60, or 90 degrees of tilt. Subjects were asked to draw a line representing the surface of the liquid in the half-filled bottles. Any line deviating from horizontal by six degrees or more was scored an error. Total scores could range from 0 to 8. Major and sex were statistically significantly related to horizontality scores, with men and science majors outperforming females and liberal arts majors. Sex accounted for less than 10% of the variance in this study. Kalichman warned that horizontality research which uses university students but fails to account for major may misrepresent actual abilities of college men and women.

It should be noted that Kalichman may have overstated the importance of academic major in explaining his data. Comparing the proportions of men and women passing the test within each college major (as reported in Kalichman, 1987) yielded effect sizes (d) ranging from 1.00 (natural science majors) to 1.28 (social science majors). In other words, although major may have influenced absolute performance on water-level tasks, the relative sex differences remained stable and large (a full standard deviation) across all five majors in his study.

There are at least two reasons to expect academic major to covary with V/H ability. First, a person's educational history may have an impact on knowledge of fundamental physical phenomena. Second, students may be expected to select majors which coincide with their individual strengths. For example, a person who has extreme difficulty comprehending verticality might find engineering to be an decidedly unpleasant field. The available evidence suggests that education is related to

V/H skills, but does not account for the usual sex difference. Observing that V/H sex differences remain stable even when academic major is controlled, Kalichman wrote, "This result suggests that academic experiences may effect task performance within and between sexes, but can not completely account for sex differences" (1987, p. 3).

Summary of Experiential Factors

Academic history, participation in athletics, and V/H training appear to be related, if indirectly, to performance on V/H tests. Differences in academic major correspond to V/H scores, although men tend to obtain higher scores than women even within majors. Athletes may be expected to obtain higher scores than non-athletes on spatial tests, although it is not clear if this pattern holds for V/H tests in particular. Training has been shown to produce statistically significant gains on V/H measures, although generalization is poor and many subjects fail to demonstrate mastery even after extensive education. It is apparent that experiential variables significantly affect V/H sex differences. But it is not clear which experiences are most important nor is it certain that environmental manipulations can lead to parity between men's and women's scores on difficult V/H tests.

Hidden Knowledge

One exciting possibility in V/H research is that people may have intact coordinate reference systems, but sometimes fail to apply their knowledge to relevant problems. Kalichman (1988) argued that there are at least six component skills involved in standard horizontality tests: visual-perceptual skills, mental rotation, image generation, disembedding, use of spatial coordinate system, and recall of relevant information. Any of these might be implicated in the sex differences observed to date. It is premature, he suggested, to presume that many women do not have intact Euclidean referent systems. He called for further component skills research. Liben and Golbeck (1986) argued that physical content in V/H tasks interferes with assessment of Euclidean awareness. They wrote, "Instructional interventions for [adult] females should be aimed at facilitating the application of an existing and potentially available conceptual framework, rather than establishing that framework de novo" (Liben & Golbeck, 1986, p. 489).

Doty (1970) found that wording of instructions can significantly influence performance on a verticality task. Asking subjects to judge whether a line was tilted or not yielded different performance than asking subjects to judge whether a line was vertical or not. It appears, then, that orthogonal and oblique are not conceptual opposites and that misapplication of oblique lines does not necessarily imply misunderstanding of orthogonality.

McGillicuddy-De Lisi et al. (1978) noted that water-line performance is not necessary to demonstrate horizontality. To test horizontality without using water, a crossbar apparatus was used. (This is the same apparatus used in De Lisi, 1983 and reviewed by Liben, 1991, all discussed above.)

First-, third-, and fifth-graders and college students (10 males and 10 females from each) were tested on both the crossbar and a water-line task. Responses were lines drawn on a pictorial representation of the apparatus (crossbar or fluid container). A short training by exposure was used. There was a significant sex difference favoring boys for fifth-graders on both tasks. But for college students the difference was statistically significant only for the water-level test.

The authors concluded that a stable concept of horizontality must have been present for accurate crossbar performance. Given the different patterns between the tasks it appears that the water-level test does not directly reflect maturity of one's coordinate axis system. A possible confound, however, is that the base of the crossbar apparatus was consistently horizontal, perhaps cuing subjects' responses. Liben's (1991) criticisms of De Lisi (1983) also apply to this research.

Blades and Spencer (1989) examined the ability of young children to use Euclidean reference systems in three experiments. In the first two experiments, children age four to six years were given two coordinates and asked to find a particular location on a grid. Some of the children had grid lines to facilitate their performance. These were not found to improve children's scores. Reference labels for vertical and horizontal axes were letters, numbers, or colored circles. Children were most accurate with colors. In a third experiment children were asked to identify the coordinates, when given the location. Most children were able to perform this task. It was concluded:

> Children's success in these experiments suggests that they are capable of combining information from two dimensions. This would indicate an appreciation of Euclidean spatial relationships by children as young as 4 years. (p. 17)

In preparation for this dissertation, much pilot research was conducted. One finding was that even people who failed water level and plumb line tasks could accurately perform other V/H tasks. For example, only one of 20 subjects made errors on multiple choice test items dealing with people standing on a hill (verticality) and a ball rolling on a sloped surface (horizontality). This suggests that some people may have accurate concepts of verticality and horizontality, but fail to apply their knowledge to all relevant situations. Of course, this pilot research was conducted informally, with no controls for confounding variables. Therefore the findings are useful primarily for guiding future research.

It has been demonstrated that people who fail water-level and plumb line tasks may successfully complete other tests of horizontal and vertical awareness (Liben & Golbeck, 1986; McGillicuddy-De Lisi et al., 1978). Furthermore, words like "vertical, horizontal," and "tilted" do not have a directly antonymous or antithetical significance for many subjects (Doty, 1970). It appears possible that most adults do have at least a rudimentary comprehension of horizontality and verticality, but often fail to apply this knowledge in appropriate situations.

Possible Mechanisms for V/H Sex Differences

Limited research has been conducted into mechanisms that might underly the V/H sex difference. But much research has been conducted investigating sex differences in spatial ability. While some have argued that V/H skills are not related to spatial ability (Petersen, 1983), others have demonstrated empirical correlations (Geiringer & Hyde, 1976) and conceptual links (Piaget & Inhelder, 1948/1956) between V/H and spatial abilities. Possible mechanisms for the sex difference in spatial abilities in general and V/H abilities in particular include biological, social, and interactional factors.

Biological Mechanisms

Biological factors that may be causally related to V/H sex differences are otolith functioning, hormonal differences, and X-linked genetics.

Sholl (1989) conducted a series of five experiments to investigate the possible correlation between horizontality and otolith functioning in college students. Although men participated in portions of the research, only data from women were used in the analysis of otolithic vestibular perception. Horizontality was measured by a water-level task involving 16 bottles of four different shapes (rectangular flask, round-bottomed flask, wine bottle, and hour-glass-shaped) oriented to octant intersections (every 45 degrees). Each orientation was used twice. Under each figure was a straight line representing a table top. Subjects were told to imagine that the bottle was sealed and half filled with water. Their job was to draw a line representing the top of the liquid. Any line deviating by more than five degrees from horizontal was considered an error. People were classified as good-horizontality subjects (GHS) if they made two or fewer errors. Those who made seven or more were termed bad-horizontality subjects (BHS).

In the first experiment good-horizontality males (GHM), good-horizontality females (GHF), and poor-horizontality females (PHF) were given the rod-and-frame test (RFT), the Purdue Spatial Visualization Test (PSVT, a measure of mental rotation skills), and the spiral tube problem (a one-item physics test). GHM did slightly better than GHF on the spiral tube problem, showing some sex difference. There was no relation between spiral tube accuracy and horizontality group for females. On the RFT an opposite pattern occurred; GHM and GHF did not differ while GHF were more accurate than PHF. The PSVT had a significant effect for sex. The PSVT horizontality effect approached significance ($p<.10$). In a regression analysis of data from the women subjects, only the RFT significantly predicted horizontality.

Because the size of the frame effect in the RFT is said to be a function of a person's relative use of vestibular and visual stimuli, Sholl decided to investigate vestibular effects on horizontality in a second experiment. Subjects from the first experiment were asked to sit in a wheel chair while they were pushed along several pathways at a predetermined speed. Their task, at the end of the short ride, was to point with a protractor device to the place where they started. To eliminate visual and auditory cues, subjects were blindfolded and listened to white noise on headphones. The Raleigh test was used to determine whether subjects tended to point the same direction (clustering) and whether they were right. GHM and GHF subjects made significantly clustered responses on some of the pathways, although only males were clustered in the target direction. PHF were random across all conditions, lending support to the hypothesis that vestibular functioning plays a role in awareness of horizontality.

Alternative explanations for the results were that PHF perceived their movement accurately, but either failed to mentally update their position or had difficulty in response production. A third experiment attempted to address these possibilities by having subjects walk rather than ride along the path. This would leave unchanged the requirements for spatial updating and pointing, but would provide kinesthetic information to supplement vestibular sensation. If subjects were more accurate under these conditions, then updating or pointing difficulties would not be viable alternative hypotheses. All subject groups showed improved performance, although PHF subjects performed randomly on two of the more difficult paths. Sholl (1989) concluded that since kinesthetic/motor efferent sensory information improved scores, vestibular insufficiency, not some other factor, must be related to horizontality scores. She wrote, "The results of Experiment 3 indicate that the sex and horizontality effects in passive transport have a vestibular basis" (Sholl, 1989, p. 119).

A fourth experiment was designed to eliminate counting as a method of judging distance and motion. Subjects were asked to repeat "the" to themselves as they rode in the wheelchair. This procedure was used to prevent subvocal articulation as a means of measuring time. Subjects were asked to determine which of two paths was longer and to indicate how much they had been turned. In this way two vestibular functions were tested: detection of angular rotation and distance displacements. There was no statistically significant effect for horizontality on angle judgments. PHF were less accurate in judging distance than were GHS. For men the ability to judge distance was correlated with judgment of angles (r=.64). For women there was no meaningful correlation (for GHF, r=.06; for PHF r=-.06).

The last experiment was designed to compare estimates of time and distance for GHF, PHF, and GHM. Time judgments did not correlate with horizontality group for women. Distance judgment accuracy did correlate with horizontality skill. GHM performed better on time judgments than GHF, but there was no significant difference on distance judgments.

The five experiments provide evidence that vestibular functioning may have an influence on horizontality among females. However, because no poor-horizontality males were included in the research, it is not clear if this factor accounts for the sex difference.

In summary, a series of experiments were conducted to test the effect of vestibular functioning on horizontality. For women college students, distance judgments which were based on vestibular sensation were correlated with horizontality. Rod and frame test accuracy was also correlated with horizontality. Judgments of time, judgments of angular motion, spatial updating, and mental rotation were not correlated with horizontality.

Research into cognitive abilities and physiological variables such as the timing of puberty and somatic androgyny suggests that sex hormones may figure into spatial abilities sex differences (Newcombe & Dubas, 1987). Sanders and Soares (1986) had college students (194 women and 80 men) take the Shepard/Metzler Mental Rotation test (three-dimensional mental rotations), the ETS Card Rotation test (two-dimensional mental rotations), and PMA Vocabulary test. All subjects were asked to rate how early they entered puberty relative to their same-sex peers using a five-point scale (much earlier, earlier, same time, later, or much later). Women were asked to report their age at menarche and men were asked to report their ages at first nocturnal emission and when they first began shaving regularly.

Ratings of puberty onset (early or late) were significantly related to scores on the Mental Rotation test (three-dimensional), but not the card rotation or vocabulary test scores. Age at menarche, shaving, or nocturnal emission were not related to any of the cognitive measures. Men had higher scores on the spatial tests, but lower scores on the vocabulary test. The authors noted that the correlation between puberty onset and spatial abilities depends on both the specific measure of spatial abilities and the way of measuring puberty onset. This study found that onset of puberty and sex accounted for unique portions of variance in spatial abilities and that the age at onset of puberty isn't the sole determinant of sex differences in spatial abilities.

Other researchers have examined the relation between age at puberty and spatial skills. The results are inconsistent. Waber (1977) found that all variance on spatial ability measures (Wechsler-Bellvue Block Design and Spatial Abilities of Primary Mental Abilities Test) that was explained by sex was also explained by age at puberty. Rierdan and Koff (1984), on the other hand, found that reported age at puberty onset was not significantly correlated with scores on the Group Embedded Figures Test (GEFT; $\underline{r}$=.02; $\underline{n}$=144) or the Digit Symbol subtest on the Wechsler Adult Intelligence Scale-Revised ($\underline{r}$=-.07; $\underline{n}$=60). Strauss and Kinsbourne (1981) found no significant correlation between reported age at menarche and score on a multiple-choice horizontality measure.

The way puberty affects spatial skills may depend on which skills are being assessed. Diamond, Carey, and Back (1983) administered the Embedded Figures Test (EFT) and a face recognition task to adolescent girls. Pubertal status was determined by height and weight measurements, pubic hair distribution, and breast development. Girls, ranging in age from 10 to 14 years, were classified as prepubescent (_n_=39), actively pubescent (_n_=40), or postpubescnt (_n_=38). They were also classified as early, middle, or late maturers. Face recognition scores were lower for actively pubescent girls relative to pre- and postpubescent girls. Relative age at puberty was not a significant factor in accounting for face recognition scores. In other words, the onset of puberty coincided with a dip in skill at recognizing faces. In contrast, pubertal status was not a significant factor in accounting for EFT scores. Regardless of pubertal status, early maturers obtained lower EFT scores than did late maturers. The authors concluded that maturation, mediated by hormones, was complexly related to spatial abilities.

Petersen (1976) assessed the relation of somatic androgyny and cognitive abilities by analyzing data from the longitudinal study conducted by the Fels Research Institute. Physical androgyny was rated from frontal photographs of nude subjects when 13, 16, and 18 years of age. Ratings were given for hair distribution, muscle development, genital or breast size, and overall body shape. Scores ranged from 1 (extremely masculine) to 9 (extremely feminine). Two cognitive measures were taken. Fluent production was the combined scores from Digit Symbol (Wechsler-Bellevue Test) and Word Fluency (Primary Mental Abilities Test). Spatial ability was a composite of scores on Block Design (Wechsler-Bellevue Test) and Space (Primary Mental Abilities Test).

To eliminate pubertal status as a confounding variable, most analyses were based only on scores for the 18-year-old subjects. Physically masculine men had higher fluency scores than spatial scores. Less physically masculine men had the opposite pattern: better scores on the spatial measures than on fluency. For women fluent production was significantly related to somatic androgyny. Masculinity, however, was positively related to women's spatial scores. In other words, physical masculinity was positively related to spatial ability in women, but negatively related to spatial ability in men. It was concluded that the results support the hypothesis that sex hormones significantly affect spatial abilities.

Some have argued that brain lateralization is related to the V/H sex difference. Waber (1977) administered a dichotic listening task to 80 children in 5th, 8th, and 10th grades. Children were asked to identify phonemes which were briefly presented in right or left ear. Lateralization was inferred from the differential scores of right- versus left-ear stimuli. Late maturers were found to be more lateralized and more successful on tests of spatial ability. It was concluded that maturational rate affects the development and organization of higher cognitive functions such as spatial ability. The interplay of lateralization, timing of hormonal surges, and learning was seen as a probable cause of sex differences in spatial abilities.

A final possible biological mechanism for V/H sex differences is genetic in nature. Thomas and Jamison (1981) proposed that gender differences in V/H performance may be due to an X-linked recessive gene. They note that within many research studies the proportion of males to females who meet criteria of V/H success often match what is predicted by X-linked genetics. Their reasoning is as follows. The phenotypic expression of X-linked recessive traits is expected to occur more often in males, at a rate equal to the square-root of the proportion found in females. So if 50% of females exhibit an X-linked recessive trait, then 71% of males should do so. In support of the argument Kalichman (1989) reported that the proportions of men and women who performed accurately on his water level tasks approximated the predictions of the X-linked recessive genetic model.

The recessive X-linked pattern has been frequently cited in accounts of gender spatial differences. Boles (1980) reviewed a large body of relevant literature and concluded, "There is essentially no convincing evidence in support of the hypothesis that a major X-linked recessive gene determines spatial ability in man" (p. 633). No integrative review specific to V/H abilities and the X-linked hypothesis was found for this dissertation. The validity of the X-linked hypothesis, as applied to V/H differences, remains unsettled.

Socialization Factors

It has been postulated that differences in socialization might underlie the V/H sex differences (Maxwell, Croake, & Biddle, 1975). Investigations of socialization effects have focused on sex role orientation, parent/child interactions, and academic history. The last two topics--parent/child interactions and academic history--have been covered earlier in this paper. The following discussion will address the relation of sex role orientation and V/H sex differences.

Kalichman (1989) investigated the relation between sex role orientation and horizontality. Subjects were 97 male and 97 female undergraduate psychology students. They were given a three-item water-level test (six degree margin of error; two or more correct counted as passing), the Self Directed Search (SDS), and Bem Sex Role Inventory. Subjects were also asked questions about their knowledge of water levels. Men performed more accurately on the water level test and more often knew the principle of invariant liquid horizontality.

There were significant sex differences which were associated with academic history. Masculine sex role scores on the Bem did not correlate with water level performance for men or women. Feminine sex role scores were negatively related to water level performance for women. For all subjects the knowledge of physical principles, measured by verbal reports, was the strongest predictor of success. Kalichman concluded:

> Performance seems to be a function of the interaction between gender roles, selection of activities and experiences, and the acquisition of physical knowledge, all of which are correlated with sex. (p. 99)

Jamison and Signorella (1980) also investigated the correlation of Bem scores and horizontality. College students (58 women, 43 men) completed the Bem Sex Role Inventory and were tested on the water-level task using the same apparatus as that in Thomas et al. (1973). Subjects' scores from the Bem were collapsed into a single score by subtracting the average rating on feminine items from the average rating on masculine items. If this score was .5 or greater the subject were classified as "masculine." If the difference was -.5 or less the subject was classified as "feminine." Subjects whose scores fell between these points were placed in the "androgynous" category. The dependent variable was number of correct responses in eight water-level trials. Any answer within five degrees of horizontal was considered correct.

Data were analyzed by log-linear categorical methods. It was found that men and women who scored "masculine" on the Bem showed no statistically significant differences on the dependent variable. Men and women subjects classified as "feminine" did considerably worse, but were not statistically different from each other. Androgynous men performed much better than androgynous women. In short, "masculine" men, "androgynous" men, and "masculine" women did well (over 70% of them passing the test). "Feminine" women, "androgynous" women, and "feminine" men did poorly (less than 40% of them passing the test). Means, standard deviations, and sample sizes for each category were not reported. The authors concluded that individuals' sex-role, independent of actual gender, plays a significant role in performance on horizontality tests.

Similar data were collected by Goodrich, Damin, and Ascione (1988). College students (59 men, 100 women) completed the Bem and were tested on a measure of verticality and horizontality. To match the procedures in Jamison and Signorella, subjects were divided into three sex-role groups. A 2 (sex) by 3 (Bem group) analysis of variance was computed. Only the main effect of sex was statistically significant. This contradicts the findings of Jamison and Signorella (1980). Thus, the effects of sex role on understanding of the Euclidean referent system remain unclear.

Cognitive Style

One correlate of V/H ability remains to be discussed. Cognitive style, or field articulation, is the tendency of an individual to perceive objects within the context of an embedding frame or to perceive the object in relative isolation from the field in which it lies. The variable of cognitive style has obvious implications for V/H performance.

Myer and Hensley (1984) had 44 women and 41 men college students take the Group Embedded Figures Test (GEFT) and a two-trial water-level task (one with an upright beaker and one tilted 45 degrees). Subjects also wrote a statement of how they guided their performance on the water-level task. Judged accuracy of statement was significantly correlated with water-level task performance, although the association was weak in practical terms. A repeated measures analysis of variance (sex by cognitive style by beaker tilt) revealed significant main effects for cognitive style and tilt and a significant interaction between the two. For sex there were no significant main effects or interactions. Two conclusions were drawn. First, the sex difference on horizontality may be mediated by cognitive style. Correcting for cognitive style eliminated the significant main effect of sex on water-level performance. Second, the importance of verbalizing the guiding principle may have previously been overemphasized as a predictor of water-level performance.

Neimark (1981) argued that cognitive style might interact with and confound performance on many measures of formal operations. She suggested that future research reduce task ambiguity by enhancing clarity, providing all information necessary for accurate performance, giving full descriptions of alternative means to the goal, and making the content realistic and familiar. Even imitation, it was argued, shows some capacity for formal operations. Pascual-Leone and Morra (1991) reviewed a broad sample of V/H literature. Regarding cognitive style they concluded:

> The overall pattern of results suggests that field dependence/independence is an important factor, but that it leaves unexplained a large proportion of variance in water level task performance. (Pascual-Leone & Morra, 1991, p. 246)

Interaction of Biology and Experience

Cramer (1971) noted that there are many more men than women in engineering. Although his arguments were originally intended to account for engineering ability, they may be no less valid when applied to spatial ability in general and V/H skill in particular. According to Cramer:

> The capacity of the penis and testicles to move and retract presents the boy with a particular challenge in the development of body image; this may contribute to his interest in machinery, physics and the like.
>
> The boy's better spatial sense relates to the greater use he makes of space in motor activity; the ability the boy has to perceive his sexual organ may also contribute to a better representation of space and to his better skill and greater interest in experimental sciences and mathematics. (Cramer, 1971. Cited in Hartston, 1987, p. 77)

Summary of Proposed Mechanisms

Many mechanisms have been proposed to explain the origin of gender differences in spatial abilities. These range from the way parents treat their sons and daughters to corporeal concerns, such as whether research subjects possess a lateralized brain, intact ears, hairy legs, and a functional penis. Taken together the proposed mechanisms illustrate the extremities to which researchers will go in their quest to explain psychological sex differences (pun borrowed from Carlson, 1980, p. 206).

Summary of V/H Research

Verticality and horizontality are empirically and conceptually related constructs. Adults make more errors on V/H measures than would be predicted by Piagetian theory or intuitive conjecturing. Men tend to obtain higher scores than women on nearly all V/H measures. Using V/H tests which are extremely easy, and thereby introducing a ceiling effect, appears to eliminate this sex difference. Numerous correlates, postulated causes, and consequences of the V/H sex difference have been advanced. Experiential factors appear to be significant, although many potentially important experiences have not been systematically examined for their relation to V/H ability. In particular, athletic experience may be expected to correlate with V/H scores. Training programs appear to increase subjects' scores on V/H measures, although generalization may be weak and some subjects may be impervious to even the most powerful interventions. It appears likely that even subjects who fail traditional V/H measures have some awareness of invariant horizontality and verticality. No training efforts documented to date have attempted to tap this hidden knowledge as a means of enhancing learning. Proposed mechanisms for the V/H sex difference include hormonal effects, brain lateralization, vestibular functioning, and X-linked genetics.

In short, it is known that adults, especially women, do not perform as well on V/H tests as might be expected. The reason for that is not clear and, as yet, no satisfactory intervention has been identified.

Hypotheses

Guided by a review of the preceeding literature it was possible to make certain predictions for the present research. The following hypotheses were tested in the present research.

Hypothesis 1: Verticality and horizontality scores are correlated. This is true for men, for women, and all subjects combined.

Hypothesis 2: Among nonathletic college students, men obtain higher scores on a measure of V/H concepts than women.

Hypothesis 3: There is no difference in V/H scores between men and women college athletes.

Hypothesis 4: Adult subjects who receive training in V/H concepts have higher scores on a V/H posttesting than do subjects who receive no such training.

Hypothesis 5: Among subjects who receive V/H training, those who are also trained to relate the new information to knowledge they already have about verticality and horizontality obtain higher V/H scores.

METHODS

Subjects

Subjects for this study were drawn from two primary sources: undergraduate psychology classes and varsity athletic programs at Utah State University. A total of 318 subjects participated. (See Table 1 for a breakdown of subject sources and number of subjects taking each test.)

Table 1

Summary of Subjects

Subject Source	N	n Men	n Women	Test Format	Sample Sizes Per Test VH1	Gn1	VH2	Gn2	VH3
Student	173	83	90	Ind	173	173	173	171	53
Students	80	31	49	Group	80	80	80	80	
Athletes	50	32	18	Group	50				
Gymnasts	15		15	Group	15				

Note. VH1 refers to the pre-training V/H test, VH2 the post-training V/H test, and VH3 the follow-up. Gn1 and Gn2 are the pre- and post-training generalization tests, respectively.

Athletes were recruited from two sources: the USU women's gymnastics team and the USU athlete study hall. Gymnasts were recruited through the assistant gymnastic coach. She acquainted herself with the research by first participating in the procedures as if she were a subject. The hypotheses were not explained to her until after the gymnasts had completed testing. Convinced that the research was interesting, meritorious, and innocuous, she took copies of consent forms and a short recruitment statement which she read to the gymnasts at a team meeting. All 15 gymnasts accepted her invitation to participate in the research even though they received no tangible reward.

Non-gymnast athletes were also recruited without inducement. The investigator arrived at the study hall and asked the athletes to volunteer five minutes of their time. Despite the fact that they would receive no extra credit or personal gain, a large majority of the athletes present at study hall agreed to participate.

A total of fifty varsity athletes (32 males and 18 females) participated. Men athletes came from the following teams: football (n=13), basketball (n=11), track (n=1), and not specified (n=7). Women athletes came from the softball team (n=7), track team (n=6), volleyball team (n=1), and not specified (n=4).

Undergraduate psychology students were recruited from introductory and developmental psychology courses at USU. They were offered extra credit or research credit for their participation. The principal investigator was invited to the class early in the academic quarter and explained to the students what the research was about, the amount of time that would be involved, and the class credit that was being offered. It should be noted that 11 varsity athletes were obtained through psychology class recruitment efforts. They were randomly assigned to groups, tested, and trained along with the psychology students. Their data were treated the same as data from nonathletes except for comparisons dealing specifically with athletics. That is, their data were included in analyses of general sex differences and effects of training.

Materials

Tests used in this research were a root beer truck test and a multiple-choice generalization test. A demographic questionnaire, consent form, and several versions of debriefing summaries were also used.

The root beer truck test, used by the author and others in previous research, was employed to assess V/H concepts. (See Appendix B for a sample of the test.) On the first of two pages is a black line drawing of a truck with a rounded tank on its bed and a pipe at the back. Subjects are guided to draw lines representing the top of the root beer (the tank is said to be half full) and a free-hanging rope (mounted on the pipe). On the second and final page are four pictures of the truck: two on level ground and two on 25 degree slants. In all pictures the truck faces to the viewer's right. The subjects' task is to draw lines representing the rope and root beer surface.

In pilot testing it was found that many subjects looked to the investigator for indications on how to draw the lines. In order to prevent inadvertent cueing, the script and intended visual cues were videotaped. (See Appendix C for the videotape script.) Copies of this tape were used for all administrations of the root beer truck test in this research, ensuring consistency across all testing environments.

There are advantages of using the root beer truck over other methods of assessing V/H concepts. The root beer truck test permits simultaneous assessment of both verticality and horizontality, can be quickly and easily administered, has an ecologically valid and simple story line that appears to render subjects at ease, and it involves plumb lines and liquid surfaces, the physical phenomena most frequently used to assess V/H concepts. Data collected by Goodrich et al. (1988) yielded the following characteristics of the test. For 552 subjects, ranging in age from 5 to 94 years (mean=18.3; median=18), the Spearman-Brown correlation between sums of errors on odd and even items is .84 (.88 for males and .81 for females). The correlation between vertical and horizontal errors ($\underline{n}$=552) is .68. The test/retest reliability ($\underline{n}$=18 college students; one week interval) is .95. It is apparent that the root beer truck test reliably assesses verticality and horizontality. Furthermore, vertical and horizontal scores are clearly related and can justifiably be summed to create a total V/H score.

The generalization test was devised to assess V/H understanding in phenomena not addressed by the root beer truck test or training procedures. (See Appendix D for a sample of this test.) The generalization test has eight items, each using a multiple-choice format. It contains four verticality items and four horizontality items. The reliability of the generalization test was assessed in two ways: split-half correlations and test-retest for subjects who received the placebo treatment. Fifty-four control subjects completed the V/H test at least two times. The Pearson correlation for their pre- and post-treatment scores was .87 ($\underline{p}<.05$). This suggests that the generalization test is reliable. A second analysis of reliability was based on split-half (odd/even) totals on the first administration of the V/H test. Subjects in this analysis were 114 men and 139 women psychology students, tested either individually or in groups (Table 1, first and second rows). The Spearman-Brown correlation between the score on odd and even items was .55, suggesting that the practical reliability of the generalization test may be more modest than indicated by test-retest figures.

Testing

Subjects recruited from athletics teams were tested in large groups. They were given a consent form, demographics questionnaire, and a copy of the root beer truck test (hereafter called the V/H test). The researcher introduced the athletes to the task by reading a script (see Appendix E). Then the videotape of the V/H test was started and the investigator stood to the side of the room. (For the gymnasts this presentation was given by the assistant gymnastics coach.) Upon completion of the test subjects were instructed to complete the demographic questionnaire. When all papers were turned in, a short debriefing was offered.

Subjects recruited from undergraduate psychology classes were divided according to sex and randomly assigned to placebo, standard training, or enhanced training groups. Assignment to groups was done after subjects signed up, but prior to actual testing. Because of subject attrition the sample sizes for these three groups are not equal.

Subjects were contacted by telephone and scheduled for testing/training. When they arrived, subjects were taken individually to a small room with a television and video tape player. There they were introduced to the task, were asked to read, sign, and date the consent form, filled out the demographic questionnaire, and completed the generalization test. (Note: 19 subjects actually were tested in the presence of one or two other people. This was done because of schedule conflicts or to reduce backlog on crowded days. In all cases the subjects were seated on oposite sides of the room and were instructed to not talk or compare their answers.) A pencil with eraser was supplied. The experimenter then gave two copies of the V/H test, two copies of a generalization test, and one demographic questionnaire. Subjects who had been previously assigned to standard or enhanced training were also given a blank 3x5 index card. Then the videotape was started and the experimenter left the room.

Tapes were prerecorded to first present the V/H test, go directly to the preassigned intervention (placebo, standard training, or enhanced training), and then conclude with a second administration of the V/H test. When the tape was finished, the experimenter entered the room, examined the materials for completeness, gave a limited debriefing, and dismissed the subject.

Within a week of the first testing an effort was made to contact subjects for follow-up testing. Follow-up testing included only the root beer truck test. Many subjects could not be contacted, had schedule conflicts, or decided to take partial credit for the portion of the research that had already completed. Others failed to arrive to their testing appointments. Consquently only 53 subjects were given the follow-up test.

Training

Training was given only to subjects recruited through psychology classes. Prior to testing, all subjects were assigned to one of three groups: enhanced training, standard training, or placebo. The placebo treatment was a Road Runner cartoon ("To Beep or Not To Beep"), 6 minutes and 30 seconds of animated physical improbabilities. In retrospect this choice of placebo may have been misguided since it involves such gross violations of physics (levitation, delayed falls, oddly contorted ballistics, etc.) that it may have actually prompted subjects to rethink their first answers on the root beer truck test.

The standard training intervention was adapted from procedures used by Barsky and Lachman (1986), Liben and Golbeck (1986, 1984), G. N. Kelly and J. T. Kelly (1978), Thomas, Jamison, and Hummel (1973), and the author's pilot study. To permit consistency in training a videotape format was used. (See Appendix F for the script.) Computer generated graphics, three-dimensional models, and trick videography were employed to illustrate the various points. Common phenomena such as lakes, people walking, falling rain, and trees on hills were shown and discussed in language carefully chosen to not sound scientific or academic. Scientific content was avoided because it had been shown to lower females' scores on some Piagetian tasks (Peskin, 1980). Duration of the standard training was 12 minutes.

Although the video training was designed to directly instruct subjects on V/H concepts, an effort was made to not teach to the root beer truck test. So trucks, liquid refreshments, and ropes were not used in the training. The generalization test was used to help evaluate whether instruction generalized to other phenomena.

At three points in the training video the subjects were asked to write something on a 3x5 card. By examining this card the experimenter could determine if the subject actually watched the training video. Only one person failed to follow these directions. Her data were discarded.

The enhanced training was identical to the standard training, but with two minutes of extra instruction. The additional training related horizontality and verticality to knowledge that subjects were presumed to already have. This presumption was based on information obtained through previous research, as explained in the "Hidden Knowledge" section of the literature review.

The enhanced portion of the training related horizontality and verticality to two phenomena that nearly all subjects seem to understand--rolling spheres and people standing. Subjects in enhanced training were taught to employ heuristic devices relating these phenomena to verticality and horizontality. For horizontality they were instructed to evaluate horizontality by asking themselves if a marble would role on a given surface. Enhanced training of verticality compared the orientation of people standing upright with oblique and horizontal surfaces. Subjects were then instructed to indicate horizontal and vertical axes by matching the angles of people's bodies and surfaces on which marbles would remain stationary.

All subjects, whether they received standard, enhanced, or placebo training, were taken individually to the testing/training room. There they were introduced to the task, asked to sign consent forms, given testing materials, and started on the videotaped presentation of the root beer truck test. The tape continued directly into the training segment and into a second root beer truck test.

Group Testing and Training

The instructor of one very large introductory psychology course offered to allow subject recruitement in his class only if all volunteers were allowed to participate. It had been anticipated that the analyses for which these subjects were being recruited would require data from approximately 80 men and 80 women. Data from more than half of those had already been collected by the time of the last recruitment drive. When more than 200 subjects volunteered from this last class, it was deemed superfluous to simply increase the sample size. It was decided to use the extra subjects in a study about one of the validity threats to the main research.

Because athletes were tested in groups, but psychology students were tested individually, differences in their scores might not be directly comparable. So the surplus subjects (80 total) were assembled in groups and given the V/H test, generalization test, and enhanced training. By comparing their scores to those of students who had been tested and trained individually, it would be possible to assess the effect of group administration and training on V/H scores. Subjects were scheduled according to convenience. All group-tested subjects received the enhanced training. (No group-tested subjects received the placebo or standard training.) Assignment to group or individual testing was done on the basis of convenience, with no effort to randomize, stratify, or otherwise control for systematic differences.

Debriefing

Upon completion of the analyses, all subjects were given a written debriefing relevant to the group from which they were recruited. The description included the purpose and findings of the research and specific results of their group. For two of the psychology classes the investigator presented the results in class. For the other classes the instructors presented the results within the context of a class discussion or lecture.

Many subjects requested a personal debriefing. The investigator personally made a telephone call to each of these people. For subjects who could not be reached on the first attempt, as many as four more attempts (each on separate days) were made. In the telephone call subjects were told about the nature of the research and the results. Many of these subjects asked if they had performed "normally." Unfortunately, the efforts to protect anonymity made it impossible to track individual performance. So the investigator made a sincere attempt to assuage any lingering doubts about academic potential, human worth, or cerebral intactness that the individuals might have. No data were collected to determine whether this effort was successful.

Scoring the Test

To score the V/H test it was placed squarely on a drafting table and secured in place. The rendered lines were then measured in whole degrees. For curved lines the orientation between line end points was used. Interrater reliability was not formally assessed. Informal comparisons of measurements made by the primary investigator and his assistant showed extremely high consistency. All lines which were within 10 degrees of correct were scored as accurate. This particular margin of error was used for two reasons. First, the test stimulus was rather small. A line which was inaccurate by even a quarter of an inch would not fall within a five degree margin. Second, the focus of this research was to investigate conceptual errors, not graphomotor problems. The 10 degree margin was deemed most useful for the circumstances.

The V/H total score was the number of lines correctly rendered on the two obliquely oriented truck pictures, with possible scores ranging from zero to four. The sum of correctly rendered rope lines constituted the V score (range 0 to 2) and sum of correctly rendered root beer lines was the H score (range 0 to 2).

Analyses

To evaluate the hypothesis of vertical and horizontal scores being correlated, Pearson product moment correlations were computed between vertical scores and horizontal scores for men, women, and all subjects combined. Unless otherwise stated, the traditional p value of .05 was used in this and all analyses.

An analysis of variance was computed for non-athlete subjects, with total scores (sum of vertical and horizontal errors) serving as dependent variable and gender serving as the independent variable. This allowed assessment of the second hypothesis, that gender differences would exist for a sample of adults.

To evaluate the third hypothesis, that there would be no gender differences in mean scores of athletes, an analysis of variance was computed for total scores of men and women varsity athletes. The dependent variable was total score on the root beer truck test and the independent variable was gender.

For hypothesis four an analysis of covariance was computed for subjects who watched a training video (standard, enhanced, or placebo). The covariate was pre-test scores. The dependent variable was posttest scores. The independent variable was whether subjects received training (standard or enhanced) or placebo.

To assess hypothesis five an analysis of covariance was computed for subjects who observed either standard or enhanced training. The independent variable was which tape they watched. The dependent variable was score on posttest and the covariate was pretest score.

RESULTS

Hypothesis 1

The first hypothesis was that verticality and horizontality scores would correlate. For all 318 subjects who participated in this research (173 undergraduate psychology students and 145 varsity athletes) V and H scores on the pre-treatment V/H test were tallied. Possible correct response scores on both V and H ranged from zero to two. The Pearson product-moment correlation between V and H was .37 ($p < .05$). (The Pearson product-moment correlation was used because the data are interval in nature and the sample sizes are large.) For men (n=146) the correlation was .27 ($p < .05$) and for women (n=172) it was .37 ($p < .05$). It should be noted that men had relatively little variance in their V and H scores. This was due in part to a ceiling effect; 64% of men obtained perfect scores on both V and H items. Only 35% of women had such high scores.

Although verticality scores do correlate with horizontality scores, the two are not equivalent. Of the 318 subjects in this analysis, 97 (31%) made errors only on either the plumb line or on the liquid surface. There were 56 people (18%) whose only mistakes were in drawing the rope. Another 41 people (13%) made errors only in drawing the liquid surface.

Hypothesis 2

It was hypothesized that men would obtain a higher mean V/H score than women. To test this an analysis of variance was computed for undergraduate psychology students taking the V/H test. (These subjects are from the first row of Table 1 in the Procedures section of this document.) The independent variable was gender. The dependent variable was number correct (range 0 to 4) on the V/H test. The mean for men (M=3.48; sd=.89; n=83) was higher than the mean for women (M=2.58; sd=1.44; n=90), $F(1,171)=24.3$, $p<.01$. Seventy percent of men obtained perfect scores. Thirty-nine percent of women obtained perfect scores. An effect size d (based on differences in means and standard deviations weighted by number of subjects) was .76 favoring men. (This method of computing effect sizes will be used throughout this document unless otherwise stated.) In other words, men performed about three-fourths of a standard deviation better than women. The second hypothesis is supported. Men obtained higher scores than women on the V/H test.

Hypothesis 3

It was predicted that there would be no difference in mean V/H scores for men and women college athletes. An analysis of variance was computed for 32 male and 33 female varsity and red-shirt athletes at USU. There was no significant main effect of gender on V/H score for this population, $F(1, 63)=2.15$, $p>.10$. For men the mean of V/H scores was 3.00 (*sd*=1.14). For women the mean was 2.55 (*sd*=1.35). A mean difference effect size *d* was .36 favoring men. The hypothesis that no statistically significant gender difference is present among athletes appeared partially substantiated. But closer examination of the data revealed a more complex pattern.

Fifteen of the women athletes were collegiate gymnasts. The other 19 women athletes participated in softball, track, field, volleyball, or basketball. The gymnasts had a mean score of 2.13 (*sd*=1.41) while non-gymnast women athletes obtained a mean of 2.89 (*sd*=1.23). The mean difference effect size between gymnasts and other women athletes was .58, with non-gymnasts performing more accurately. In fact, while 7 of 19 (37%) non-gymnasts obtained perfect V/H scores, only 3 of 15 (20%) gymnasts did as well. Why gymnasts would perform so poorly is unclear.

A similar breakdown was possible for men athletes if grouped as football (*n*=13) or non-football (*n*=19). Mean V/H scores for the two groups were identical. No main effect of sport (at least given the only possible sport breakdown available) is present for men.

Although the hypothesis of no gender differences in V/H scores among athletes is ostensibly supported by these data, the relation of athletics to V/H understanding appears to be considerably more complex than originally anticipated.

Hypothesis 4

The fourth hypothesis was that subjects who received training would obtain higher mean scores on the V/H test than subjects who received no such training. There were 24 men and 55 women who made at least one error on the initial V/H test (hereafter called V/H-naive subjects) and who then watched either a training tape or the placebo tape. An analysis of covariance was performed on data from these V/H-naive subjects, with first V/H score as the covariate, training tape or placebo as the independent variable, and score on a second administration of the V/H test as the dependent variable. There was a significant effect for treatment on V/H score, $F(1, 76)=6.72$, $p<.05$. V/H-naive subjects who observed a training video had a higher V/H mean on retest ($M=3.09$; $sd=1.39$; $n=53$) than V/H-naive subjects who watched a placebo tape ($M=2.50$; $sd=1.24$; $n=26$). The effect size difference between the two groups, using the effect size at posttest minus effect size at pretest, is .69 favoring subjects who watched a training tape. The hypothesis appears substantiated; subjects who received V/H training obtained higher retest scores than did subjects who received no such training.

It was possible to give the V/H test a third time to 53 of the above subjects. The other 26 subjects failed to arrive for follow-up testing or could not be contacted to schedule testing. Drop-out rates were similar for the treatment and control subjects: 32% and 35%, respectively. The median interval between initial and follow-up testing was 17 days (mean=16.7 days; sd=10.6). An analysis of covariance was computed using follow-up scores as the dependent variable, pretest scores as the covariate, and treatment/placebo as independent variable. There was no significant effect of training/placebo on mean scores at follow up, $F(1, 50)=1.56$, $p>.10$. The effect size difference between the two groups, using effect size at pretest minus effect size at follow-up, is .56 favoring subjects who watched a training tape. Examination of the means and standard deviations in Table 2 shows that treatment subjects improved their scores with testing and maintained that improvement. Placebo subjects made smaller gains after viewing the cartoon, but showed continued improvement for the follow-up test. It appears that the gains made by subjects after watching a training video are not short lived. However, repeated testing may improve scores, as shown by the linear rise in means for the placebo group.

Hypothesis 5

It was hypothesized that subjects who received enhanced training would obtain higher mean V/H scores than would subjects who received standard training. An analysis of covariance was computed for V/H-naive subjects who received training. The independent variable was whether subjects viewed the standard or enhanced training. The dependent variable was score on the post-training V/H test. The covariate was pre-training V/H score. There was no significant difference between means of those who observed the enhanced training tape and those who observed the standard training tape, $\underline{F}(1, 50)=.61$, $\underline{p}>.25$. Means and standard deviations for the enhanced and standard treatment groups on the pretest, posttest, and follow-up test are presented in Table 3. An effect size, computed by subtracting the effect size at pretest from the effect size at posttest, was -.36 favoring those who observed the standard training.

Table 2

Mean V/H Scores for All Naive Subjects

V/H-Naive Subjects: Training Versus Placebo									
Group	----Test 1----			----Test 2----			----Test 3----		
	Mean	SD	n	Mean	SD	n	Mean	SD	n
Training	1.77	1.05	53	3.09	1.39	53	3.06	1.45	36
Placebo	2.04	1.08	26	2.50	1.24	26	2.71	1.26	17

Analysis of covariance was also conducted for the follow-up test, using pretest scores as the covariate, treatment group as the independent variable, and follow-up scores as the dependent variable. There was no difference between mean scores on the follow-up test for the two groups, $F(1,33)=.33$, $p>.25$). The effect size, computed as above, was .11 favoring subjects in the enhanced training. The hypothesis that the enhanced training video would lead to higher scores on V/H testing was not supported. Training that relates difficult aspects of V/H knowledge to readily understood V/H phenomena, at least as it was conducted in this research, did not appear to be superior to the traditional training methods discussed in the literature review.

Table 3

Mean V/H Scores for Naive Subjects

V/H-Naive Subjects Who Received Training									
Training	----Test 1----			----Test 2----			----Test 3----		
Group	Mean	SD	n	Mean	SD	n	Mean	SD	n
Enhanced	1.83	1.07	29	3.00	1.51	29	3.20	1.40	20
Standard	1.71	1.04	24	3.21	1.25	24	2.88	1.54	16

Supplemental Analyses

The results described above pose many additional questions. Some of the questions lend themselves to evaluation through analyses that were not part of the original research plan. These additional analyses will be considered in this section.

The Relation Between V/H Knowledge and Athletics

Group Versus Individual Testing

Athletes in this study were tested in groups, whereas other subjects were tested individually. The testing format may have affected the scores obtained by these two groups. To determine if administration format (individual or group) influenced scores, the V/H test was administered to 80 additional psychology students in groups of 2 to 10 people (Table 1, row 2). Their V/H scores can be compared to the V/H scores of subjects who were tested individually (Table 1, row 1). Because 19 of the original subjects were tested in the presence of one or two other people, the sample sizes are as follows; 99 subjects were tested in groups and 154 subjects were tested individually.

An analysis of variance was computed using V/H score on the pretest as the dependent variable and group/individual administration as the independent variable. Testing format had no effect on means of V/H score, $F(1, 251)=.27$, $p>.25$. Means were 2.95 ($sd=1.31$; $n=154$) for subjects tested individually and 3.04 ($sd=1.23$; $n=99$) for those tested in groups. The effect size for these two groups was .05 favoring group testing. It appears that whether testing was conducted individually or in groups did not significantly influence V/H scores.

It is possible that men and women are differentially influenced by the presence of other people. To test this, analyses of variance were calculated separately for men and women. Again there was no effect of testing format on mean V/H scores. For men the F was 1.27 (df=1,112; $p > .25$) and for women the F was 1.82 (df=1, 137; $p > .10$). (See Table 4 for means and standard deviations across groups.) It did not appear to matter substantially whether group or individual testing was used.

Athletics and V/H Scores

All subjects were asked to rate their current level of sport activity in comparison with other people their age. A one to five scale was used, with five being frequent participation and one being infrequent participation. Using data from all subjects who responded to the sports rating item (n=317) a Pearson product-moment correlation between the rating and the V/H score was near zero ($r = -.03$; $p > .25$). Separate correlations were computed for men and for women. Women's V/H scores did not correlate with sports participation ratings ($r = -.04$; n=172; $p > .25$). For men, however, the ratings did correlate ($r = -.25$; n=145; $p < .01$), with frequent sports participation corresponding to lower scores on the V/H test. As was the case in the analyses for hypothesis four, the relation between athletics and V/H knowledge appears to be complex.

Error Analysis

What Types of Errors were Made on the Root Beer Truck Test?

There have been some attempts in previous research to classify or analyze error types. Harris et al. (1978) reported that most horizontality errors made by college students were at 10 degrees, the smallest foil offered in their multiple choice water level test. They also noted that in less than five percent of responses did subjects select an answer that was parallel to the tilt of the figure. Wittig and Allen (1984) also classified error types and reported the proportion of each error relative to the total number of errors.

Table 4

Mean V/H Scores as a Function of Testing Format

	Men			Women		
Format	Mean	SD	N	Mean	SD	N
Individual	3.52	.83	69	2.49	1.44	85
Group	3.31	1.16	45	2.81	1.25	54

A more extensive consideration of water level response types was presented by Pascual-Leone and Morra (1991). They described four types of responses and analyzed their distribution within five sets of data from previous research. Their response classification system followed from the model formulated by Pascual-Leone and has been applied only to horizontality data, although conceivably it might also apply to verticality measures. Their categories were: accurate responses (within five or ten degrees of horizontal), compromise responses (moderate errors of less than 30 degrees), bottom driven responses (essentially parallel to the bottom of the glass), and excessive responses (deviating by at least 50 degrees). Pascual-Leone and Morra (1991) found the system to be useful for evaluating the water level responses made by children and adults.

There were characteristics of the Pascual-Leone/Morra response classification system that made it of limited use in the present research. First, it considered only absolute deviation from horizontal, disregarding the orientation of the test stimulus. In this way lines sloping in opposite directions were grouped together. Second, the "compromise" category engulfed 90 degrees of orientation, fully one-half of the total range. Third, one type of responses was rarely made by adults. Fourth, the system was tied to the complex cognitive model promoted by Pascual-Leone. Its generalizability may have thus been limited. For the present research a different classification scheme was needed and devised.

Errors in V/H research can be classified as undercorrections, overcorrections, or miscorrections, depending on their orientation relative to the slope of the oblique test stimulus. (See Appendix G for an illustration of how these errors were classified.) In this research all liquid surface lines that were drawn nearly parallel to the sloped ground or rope lines that were approximately perpendicular to the sloped ground were labelled undercorrections. Specifically, lines sloping at least 11 degrees in the orientation of the hill (clockwise or counterclockwise) and not more than 30 degrees were labelled undercorrections. (Recall that the hills in the V/H test slope -25 and +25 degrees.) In making this type of error a person responds as if the ground were level, not sloped. This is the most common type of error made by young children. A reanalysis of data collected by Goodrich et al. (1988) showed that all errors on the V/H test made by 48 first grade students were undercorrections.

Overcorrections are lines which go too far in compensating for the slope of the ground. If the ground slopes upward to the right, an overcorrected liquid line will slope up to the left. Any line deviating by at least 11 degrees in this fashion was labelled an overcorrection. A person who makes this type of error shows awareness that the rope and liquid don't slope with the hill and responds as if the correct orientation were opposite the slope of the hill.

Miscorrections are lines which exaggerate the slope of the distracting stimuli. So if the ground in a test item slopes by 25 degrees, a miscorrected liquid line is angled more than 25 degrees. For this research all lines which exaggerated the slope of the hill by at least six degrees were labelled miscorrections.

What Patterns are Apparent in the Errors?

For the 318 subjects in this research, their pre-training V/H tests produced 937 correct responses, 223 overcorrections, 57 miscorrections, and 55 undercorrections. Of these 318 subjects 163 subjects made at least one error on the pre-training administration of the V/H test. Among these 163 V/H-naive subjects, overcorrections were the most common type of error, with 128 subjects (79%) making at least one such response. Miscorrections were made by 39 (24%) people. Thirty-four individuals (21%) made at least one undercorrection. Because some people made more than one type of error, the percentages have a sum greater than 100.

Recall that subjects were asked to draw four lines near oblique surfaces. They could therefore make as many as four errors. Of the V/H-naive subjects, 36 (22%) made more than one type of mistake on the first administration of the V/H test. Most of these people made at least one overcorrection and at least one miscorrection ($\underline{n}$=23; 14% of all V/H-naive subjects). Of the 36 subjects who made more than one type of error, all but three made at least one miscorrection.

There is another way to look at error combinations. Of the subjects who made undercorrections 38% also made other errors. Among overcorrectors 22% made other kinds of errors. But for those who miscorrected 85% made other types of errors. It appears that miscorrection corresponds to random responding, with a "pure guess" strategy uninformed by any particular algorithm.

If miscorrectors are performing at random, they might be expected to make relatively numerous errors. For V/H-naive subjects who made undercorrections or overcorrections the mean V/H score was 1.92 (<u>sd</u>=1.04, <u>n</u>=157). Among miscorrectors the mean was 1.20 (<u>sd</u>=1.00, <u>n</u>=39). This yields an effect size of .70. In other words miscorrectors scored two-thirds of a standard deviation worse than other V/H-naive subjects. (Because some subjects were in both groups, no tests for statistical significance could be performed on these data. Although it is possible to dichotomize subjects as miscorrectors or nonmiscorrectors, an artifactual difference in scores results. On any given line miscorrectors could make a correct response or one of <u>three</u> error types, whereas nonmiscorrectors could make a correct response or one of <u>two</u> error types.)

Does the Type of Error Make a Difference on How Subjects Respond to Training?

There were 96 subjects who made errors and received some training, either standard or enhanced. Because testing format (individual or group) had no effect on pre- or post-training V/H scores, this sample included subjects tested individually and subjects tested in groups.

Of these 96 V/H-naive subjects who received training, 26 made miscorrections on the first administration of the V/H test and 70 made only other types of errors. Of the miscorrectors, 19 (73%) were women. Of other V/H-naive subjects, 51 (73%) were women. An analysis of variance was computed using post-training V/H scores as the dependent variable and type of error (miscorrection or other) as the independent variable. There was a significant effect of training on posttest mean scores, $F(1, 94)=5.63$, $p<.05$. Miscorrectors had a mean score of 2.81 (sd=1.52), while other V/H-naive subjects obtained a mean of 3.46 (sd=1.05). The effect size is .55, favoring subjects who did not make a miscorrection on the first V/H test.

How do Miscorrectors Compare to Other Naive Subjects When Given a Placebo?

An analysis of variance was computed, using error type on the first V/H test (miscorrection or other) as the independent variable and score on the second V/H test as the dependent variable. Subjects were 26 V/H-naive subjects who watched the placebo tape. The mean post-training scores of the two groups were not significantly different ($F(1, 24)=1.03$; $p>.25$). For miscorrectors the mean was 2.00 ($sd=1.41$; $n=6$). Other V/H-naive subjects had a mean of 2.60 ($sd=1.23$; $n=20$). The effect size was .47, favoring those who did not miscorrect on the first test.

How Much Post-Treatment Variance is Explained by Error Type on Pre-Treatment Testing?

Product moment-correlations were computed between scores on the second V/H test and the dichotomous variables of gender, training (yes or no), and miscorrection (yes or no). Subjects were 172 psychology students tested individually and randomly assigned to one of three videos (enhanced training, standard training, or placebo). Correlation coefficients were .11 ($p>.05$) for training (trained subjects did better), .20 ($p<.05$) for gender (men did better than women) and .32 ($p<.05$) for error type (miscorrectors did worse than others). In other words, error type accounted for twice as much variance as gender and eight times as much variance as training on scores of the second V/H test.

To summarize the findings regarding error types, subjects who miscorrect on the V/H test make many errors, show a variable response pattern, and gain relatively little from training or placebo. The variable of error type is more powerful than the variables of gender or training in predicting subsequent scores on the V/H test.

Generalization

How do Men and Women Compare on the Generalization Test?

An analysis of variance was computed using the generalization test score as the dependent variable and gender as the independent variable. Subjects were the same 253 students in the previous analysis, all of whom took both the V/H test (with either group or individual administration) and the generalization test. For men the mean score on the generalization test was 7.20 (sd=1.03; n=114) and for women it was 6.38 (sd=1.47; n=139), $F(1, 251)=25.25$, $p<.01$. The effect size was .65, favoring men.

Does Training Improve Performance on the Generalization Test?

To answer this question an analysis of covariance was computed, using pre-treatment generalization test score as the covariate, training/no-training as the independent variable, and post-treatment score as the dependent variable. Subjects were the same as in the two previous analyses, with the exception of two individuals who did not complete the second generalization test. Training has a significant effect on subsequent generalization test scores ($F(1, 248)=6.10$; $p<.025$). Trained subjects obtained a mean post-training generalization score of 7.21 ($sd=.91$; $n=197$) and placebo subjects had a mean of 6.98 ($sd=1.12$; $n=54$). The effect size is .24, favoring subjects who received training. It appears that training improves V/H performance on a variety of measures, not just the V/H test designed for this research.

Did it Matter Whether Subjects Viewed Standard or Enhanced Training?

An analysis of covariance was computed for subjects who received training, with training tape (enhanced or standard) as the independent variable, score on the second generalization test as dependent variable, and first generalization test score as covariate. Subjects were the same as in the previous analysis. There was no difference between the generalization test mean for those who watched the enhanced training ($M=7.17$, $sd=.94$; $n=145$) and the mean for those who observed the standard training tape ($M=7.33$; $sd=.81$; $n=52$), $F(1, 195)=1.20$, $p<.25$. The effect size was -.16, favoring subjects who observed the standard training tape. The specific training tape that subjects viewed did not significantly affect post-treatment scores.

DISCUSSION

Previous investigations of V/H knowledge have consistently yielded three findings. First, many adults perform more poorly than predicted by Piagetian theory. Second, men tend to obtain higher V/H scores than women. Third, training programs are moderately beneficial, although a contingent of subjects is quite resistant to even intensive efforts to teach accurate V/H awareness. These studies have been so extensive and compelling that any V/H research which finds near-perfect V/H performance in adults, no V/H gender difference, or ineffective training could be dismissed as flawed or anomalous.

In this study many subjects, especially women, made errors on the V/H test. The magnitude and direction of the gender difference matched that of previous research. Additionally, the training interventions utilized in this study were only modestly successful. All of these findings match well with previously established facts. Given that this research utilized a novel V/H instrument, the fit with previous research is critical.

By themselves the above findings offer little new information about the implications, causes, or remedies for poor V/H judgment. But there were many findings original to this study which do provide such information. The following discussion will focus on the findings that extend what was already known about V/H awareness.

As with much research, the most intriguing results of this study were not anticipated. Two surprises stand out. First, there is a type of error on the root beer truck test that seems to signal extreme Euclidean naivete. Second, women gymnasts have lower V/H scores than all other adult groups studied.

Also common to most research, this study produced some disappointments. First, participation in athletics does not seem to be directly related to V/H success. Second, this research produced no direct evidence of "hidden knowledge" in adults who make errors on a V/H test.

Miscorrection Errors

A significant finding of this study was that miscorrections are fundamentally different from other responses on the V/H test. This conclusion was reached through several different empirical analyses and inspection of inferred cognitive strategies associated with the four V/H responses.

Empirically, miscorrectors were shown to be different from other subjects. They had lower V/H scores than other V/H-naive subjects at the pre-intervention test, after viewing a placebo, and after receiving training. In short, miscorrectors had consistently lower scores than other V/H-naive subjects.

One of the most distinguishing characteristics of miscorrection was that other errors were so often present on the same test page. More than one type of error was made on the V/H test by 38% of undercorrectors and 22% of overcorrectors. But 85% of miscorrectors made at least two types of errors. The most common combination was miscorrection with overcorrection. Consider the appearance of a miscorrection and an overcorrection together on a V/H test. (See Appendix H for an example.) This particular combination of errors produces a situation in which the plumb line becomes approximately parallel to the liquid surface.

Mature Euclidean concepts have two features: perceiving the world through a perpendicular grid and gravity orientation of that grid. All subjects who drew at least one line more than 10 degrees beyond true horizontal or vertical failed to demonstrate dependable gravity orientation. But only miscorrectors failed to maintain perpendicularity. Whereas all V/H-naive subjects portrayed inaccurate V/H lines, only miscorrectors depicted the rope and liquid surface in a way that substantially violated their mutually perpendicular nature. This is another way in which miscorrectors were distinct from other subjects.

Miscorrectors are different from other subjects in the strategies they employed on the V/H test. To illustrate this point consider the thinking patterns that might produce the four responses possible in this research: undercorrection, overcorrection, miscorrection, and correct response. (See Appendix I for a qualitative analysis of these solution strategies.)

The least frequent response in this study was undercorrection. Undercorrections will result from heuristics such as "make the liquid even with the ground" and "draw the rope toward the ground." These methods rely on the ground and will produce accurate results only when the setting is level. Undercorrections were the least common response in this study and only three adults made this error on all four oblique test stimuli on the pre-intervention V/H test. Undercorrection represents a systematic, ground-based solution strategy.

Overcorrection was the most frequent type of error in this research. Overcorrectors exaggerate the visual discrepancy between sloped ground and gravity-defined orthogonality. A heuristic that would lead to overcorrection is, "If the ground is sloped one way, then draw the liquid sloping the other way." Recognizing and inverting a slope requires reliance on a stable horizontal or vertical referent which is independent of the ground, a characteristic not present in undercorrections. Overcorrection represents a common, systematic strategy that combines ground- and gravity-based information.

Accurate responses were the most frequent type of response in this study. Correct responses must derive from strategies that involve gravity or other external, stable referent. Two of the gravity-based algorithms that would result in consistently accurate V/H lines are, "Draw the top of the liquid perpendicular to the pull of gravity and draw the rope parallel to the pull of gravity." Correct responses represent common, systematic, gravity-based solution strategies.

It is difficult to imagine any systematic mental strategy that would lead to errors of miscorrection. Perhaps some insight into the relevant thought processes can be gained by considering the experience of a psychology graduate student who took the root beer truck test as a favor to the investigator. Fred (a pseudonym) was baffled by the test. On the liquid surface of the uphill truck he made an undercorrection, erased, overcorrected, began laughing, erased, miscorrected, and started joking and rationalizing his performance. His discomfort spiralled upward until the investigator felt obliged to terminate the test and debrief. By that time the test page had been erased so often that the paper nearly had a hole in it.

This informal case study illustrates what some miscorrectors may experience. Fred approached the task with a ground-based strategy, disregarding the slope of the hill. Upon visual analysis Fred realized that the slope invalidated his simplistic, ground-matching strategy. He countered with an overcorrection, but recognized it as another error. His efforts quickly degenerated into unabashed guessing. Twice he rendered and erased lines which were nearly horizontal! Unlike all other responses, miscorrections are the only response which stem from unsystematic, groundless guessing.

A number of error analyses were conducted for this study and all led to the same conclusion; miscorrections represent a fundamentally different response from all other V/H responses. Perhaps the distinct nature of miscorrections can account for some perplexing results in previous research.

Relation to Previous Research

Liben (1978) noted there are two hypotheses to account for V/H failures in adults. The first hypothesis is that the subjects have a competence deficit and really don't know about verticality or horizontality. The second hypothesis is that the adults do know about verticality and horizontality, but sometimes fail to apply their knowledge, a performance deficit. There have been several studies which attempted to resolve which hypothesis better accounts for adult failures in V/H testing (De Lisi, 1983; Liben & Golbeck, 1984; Liben, 1978; Golbeck, 1986). These studies have been inconclusive, the data offering partial support for both hypotheses.

A second research strategy that has netted mixed results is training. Most training efforts show significant V/H gains for many subjects, but a stable contingent of immutably naive subjects fails to "get it" even after creative instruction, discovery experiences, and physics lectures (G. N. Kelly & J. T. Kelly, 1978; Liben, 1978; Liben & Golbeck, 1984; Thomas et al., 1973).

It has been contended here that miscorrectors are substantially different from other subjects on V/H tests. If the claim proves to be true, it might help explain the two sets of perplexing research. Perhaps miscorrectors, the only subjects who violate both orthogonality and gravity orientation of mature Euclidean concepts, are truly V/H-naive. This would suggest that the performance deficit hypothesis accurately accounts for most adults who make V/H errors, but that competence deficit accounts for the performance of miscorrectors. It would also help explain mixed training results. Overcorrectors and undercorrectors would be expected to demonstrate gains in V/H knowledge, while miscorrectors should derive little benefit.

Treating miscorrectors as members of a distinct population serves as a method of organizing data. It does not explain why there are sex differences or why so many adults misapprehend V/H phenomena. The ultimate question of cause remains unanswered by the error analyses.

Implications of Error Types

There are several implications of the observation that error type constitutes a significant variable in V/H research. First, further research into causes, treatments, and distributions of V/H naivete may be more efficiently focused if error types are considered. For example, the competence/performance debate may never be resolved until it is acknowledged that both hypotheses may be correct within limited groups of individuals.

Second, any efforts to use multiple-choice V/H test formats should include all response possibilities. Otherwise miscorrection errors may be missed.

Third, there may be much extant data which could be reanalyzed to either support or refute these findings. This would permit quick progress in determining whether error analysis represents a truly significant advance in the study of Euclidean awareness.

Finally, some people--miscorrectors in particular--may do well to avoid V/H-related fields such as civil, mechanical, or aerospace engineering, architecture, and construction. Until more efficient training programs are available, miscorrectors might find the requirement for unerring V/H awareness to be oppressively difficult. But such awareness is at times critical in more than an academic sense. This author finds it frightening to think that dams, highways, buildings, and airplanes might be designed or maintained by people who believe plumb lines and liquid surfaces are sometimes parallel.

Threats to Validity

Analyses of score distributions, error combinations, training effects, placebo effects, assessment of solution strategies, and fit with previous research all suggest that error type is an important variable. The fact that this conclusion can be reached from several types of analysis grants it the power of strong inference. There are, however, certain threats to validity which must be considered.

The first question is whether the root beer truck test is valid. The test has content validity in that it rests on two phenomena which are clearly relevant to V/H knowledge and to previous research. Although there is no authoritative list of V/H phenomena, it seems safe to conclude that there are many other phenomena which were ignored in this research. The root beer truck test does, however, tap natural phenomena which have been used extensively in previous research. In this respect the root beer truck test has content validity.

Concurrent validity was assessed in this study by the generalization test with which the root beer truck test correlated at a statistically and practically significant level. Discriminant validity was not an intended feature of the root beer truck test, but was demonstrated by its capacity to predict which subjects would benefit from training and which would not. One type of construct validity has also been established: the root beer truck test produces results which closely match well established phenomena. As described at the beginning of the Discussion section, the present research has found that many adults, especially women, make V/H errors and that training is only partially successful in remediating V/H naivete. In this way the root beer truck test is a valid instrument for investigating these phenomena.

Random assignment to treatment, use of a placebo group, and videotaped test administration eliminated many of the more common threats to internal and external validity. The different formats for testing (group or individual) had no main effect on mean scores for men, women, or all subjects together and therefore pose little concern for validity. Although there was no effort to randomly select subjects, the bulk of research on V/H concepts, like the present study, has dealt with college students. Although this might conceivably pose a generalization threat, the danger seems rather remote.

Low Scores by Women Gymnasts

A second surprise in this study was the relatively poor performance by women gymnasts. The gymnasts had the lowest mean V/H score of all groups in this study, including men and women in undergraduate psychology classes, men in varsity athletics, and nongymnast women varsity athletes.

Threats to Validity

There is one major threat to the internal validity of the research comparing gymnasts to other college students. Gymnasts were recruited for this study, introduced to the test, and monitored during the test by their assistant coach. It is possible that this somehow created a meaningful change in the V/H test. But gymnasts received the same videotaped test administration as that given to all other subjects. Any confounding variable would have to be both subtle and powerful to create such a score discrepancy in otherwise identical testing circumstances.

External validity suffered a threat due to the small (n=15) and selected sample of gymnasts. It is clearly too early to conclude that the population of collegiate women gymnasts obtains relatively low scores on tests of V/H skills. The finding is nonetheless intriguing and merits further study.

There is at least one mention in the literature of a finding similar to the relatively poor performance by gymnasts. Olson et al. (1988) found that many sporting and other activities correlated positively with spatial scores. But for women there was a significant negative correlation between a composite spatial score and participation in ballet and choreographing dance. Women's gymnastics is permeated by ballet and dance choreography and may be seen as a closely related activity. In this way the present finding has a precedent.

Relation to Previous Research

Four lines of research may be related to the findings involving gymnasts: somatic androgyny, vestibular acuity, the neuropsychological distinction between procedural and declarative knowledge, and the sensory modality used in solving the problem.

Petersen (1976) reviewed data from the Fels Research Institute for the Study of Human Development, focusing specifically on data related to physical androgyny and cognitive measures. She found that spatial abilities (measured by Block Designs of the Wechsler-Bellevue Test and Space from the Primary Mental Abilities Test) correlated negatively with ratings of physical femininity (measured by ratings of proportions of muscle to fat, overall shape, breast size, and pubic hair distribution) for women 13 to 18 years of age.

In a partial replication, Berenbaum and Resnick (1982) combined data from four longitudinal growth and development studies. They found a similar pattern of androgyny and cognitive skills, although the overall effect was smaller than that found by Petersen.

The above two studies provide limited evidence that physically masculine women have a slight spatial advantage relative to physically feminine women. It might be argued that gymnasts constitute a population of athletic but physically feminine women and might therefore be expected to have relatively low scores on spatial tests.

The relation of vestibular perception and V/H ability was studied by Sholl (1989) who found a positive correlation between the two variables. She concluded that poor vestibular acuity might interfere with the observation of V/H phenomena in oblique surroundings and might consequently prevent development of accurate V/H awareness.

The present finding that gymnasts obtain low V/H scores argues against Sholl's conclusion. If V/H maturity required sound vestibular functioning, then a balance-intensive sport like gymnastics would screen out all vestibularly handicapped individuals. The population of collegiate gymnasts would then be comprised of individuals who do not have the vestibular risk factor which purportedly accounts for low V/H scores. Gymnasts would be expected to have higher, not lower V/H scores. The data collected here refute the hypothesis that vestibular functioning accounts for V/H inaccuracies.

A third line of related research is found in neuropsychological literature. A discontinuity in performance and declarative expressions of V/H knowledge might even be predicted. Mandler (1988) noted that there is a distinction "between knowing how to see and how to move through space...and knowing that certain spatial relationships obtain in a given situation" (p. 424).

One case study provides compelling evidence that declarative V/H knowledge (e.g., taking a V/H test) does not always match procedural knowledge (e.g., performing on a balance beam). Horizontality, verticality, and other orientation discriminations were assessed for a 36-year-old woman, DF, who had suffered brain damage from carbon monoxide poisoning. She exhibited severe visual form agnosia, with extremely poor ability to recognize orientation and shape. For example, when shown a vertical block she judged it to be horizontal. However, two expressions of orientation perception remained intact for DF: visuomotor guidance tasks and the McCollough effect. The visuomotor aspects were described in Goodale, Milner, Jakobson, and Carey (1991). They noted that when DF reached for rectangular blocks she was observed to correctly orient her hand (horizontally, vertically, or oblique) prior to touching.

The McCollough effect is a visual aftereffect based on color and orientation of repeatedly presented lines. Over the course of several minutes the subject is shown alternating patterns. A typical procedure might involve one pattern with vertical green and black lines and another pattern with horizontal red and black. Because of sensory adaptation the subject is likely to perceive subsequent horizontal white and black lines as green, while vertical white and black lines are seen as red. The effect is specific to orientation, size, visual field area, and eye of induction.

Humphrey, Goodale, and Gurnsey (1991) reported on DF, the same patient described above, focusing on her experience of the McCollough effect. Despite DF's visual deficits she was able to accurately identify colors. DF was shown alternating colored adaptation stimuli for 10 minutes, then presented with black and white grids. The McCollough effect was present; when shown vertical black and white bars she reported seeing a reddish figure. When a vertical test pattern was slowly rotated, the red was perceived to fade until, at 45 degrees, the stimulus was perceived as achromatic. As the stimulus was rotated toward horizontal she began to perceive green, the complementary color for the horizontal adaptation stimulus.

Implications of this case study are that, at a neurological level, perception of orientation is not a unitary phenomenon. Physical performance and perceptual aftereffects may demonstrate V/H perception which is not matched in declarative knowledge. This helps explain how gymnasts, with keen visuomotor expressions of orientation sensitivity, might obtain the lowest mean V/H scores of any group tested for the present study.

The fourth line of related research comes from unpublished research involving gymnasts. R. Gordin Jr. (personal communication, February 25, 1992) described fascinating research involving gymnasts and a paper and pencil test that had reportedly been used by the East German national gymnastic team. The test presented five sequential drawings of common gymnastics moves. The fourth frame was left blank. Test subjects were asked to fill in the missing picture. American collegiate gymnasts performed badly on this task, whereas nongymnasts who were involved with gymnastics (for example, coaches and trainers) found the test unchallenging.

One interpretation of these data is that proficient gymnasts become so attuned to kinesthetic cues that they become relatively unskilled at visual problem solving. In contrast to the popular perception of gymnasts, mental preparation for routines is not based on visualization, but rather on a kinesthetic form of imagery. Indeed, attempts to visualize often result in performance decrements.

The implication is that balanced reliance on both kinesthetic and visual cues is most likely to result in high V/H scores. This interpretation helps reconcile Sholl's (1989) findings with the low mean V/H scores by gymnasts in the present study.

The Relation of Athletics to V/H Performance

"The great tragedy of science--the slaying of a beautiful hypothesis by an ugly fact."--T. H. Huxley (quoted in Rawson & Miner, 1986, p. 265).

As was hypothesized, there was no significant difference in mean V/H scores between men and women varsity athletes. It had been anticipated that sports would serve to screen out subjects with poor V/H skills and strengthen V/H awareness in those who participated extensively. Gender differences were not foreseen because both men and women athletes were expected to perform at or near the test ceiling. Although the data supported the hypothesis, they did not match the underlying rationale. The correlation between rated sports participation and V/H scores was negligible. There was no effect of varsity athlete status on mean V/H score. In this particular case, then, the hypothesis received superficial support while being soundly rejected at a more profound level.

Hidden V/H Knowledge

This research produced only indirect evidence of "hidden knowledge" in adults who make errors on a V/H test. The direct test pivoted on the difference between standard and enhanced training videos. The differences between scores of subjects in the two training programs were well within the range expected by random fluctuation.

For those who balk at the idea that so many adults may actually not know which way gravity pulls, there is indirect evidence to buttress their opinion. The entire argument about the significance of error type coincides well with the belief that most adults must be capable of level thinking. But this support comes with a price tag; to argue that most V/H errors do not represent competence deficits it must be agreed that some people really don't know which way is up.

Future Directions

It is impossible to assess perpendicularity of people's concepts of vertical and horizontal phenomena if both axes are not measured. Yet much V/H research assesses only one or the other, what might be termed the half-axis approach. It appears from this study that perpendicularity may be a critical factor in determining which subjects will eventually demonstrate sound Euclidean concepts. Future research may be improved by including measures of both verticality and horizontality.

This research clearly indicates the need for two immediate research tasks. First, analysis of error types should be applied to other samples and V/H instruments. Second, another group of gymnasts should be compared to a matched sample to determine if gymnastics is consistently negatively related to V/H awareness. The explanatory power of this would be enhanced if "think-aloud" testing were employed. This might help clarify the role of sensory modality (visual, kinesthetic, or both) in V/H problem solving.

Other research that may be appropriate includes further efforts to delineate characteristics of effective V/H training and factors relevant to generalization of training.

A final research area that appeals to this investigator despite the lack of empirical support is the search for V/H phenomena which are accurately understood by nearly all adults. The distinction between the enhanced and standard training interventions in this study rested on the assumption that the vast majority of people would show more accurate V/H concepts when tested by questions about rolling balls and standing people than when measured with liquid surfaces and plumb lines. There is no empirical backing for the assumption, but it has such intuitive appeal that this investigator finds the research extremely inviting.

Summary

It is disconcerting to consider how many of today's college students literally do not know which way is up. This research has not explained why it is so, but has opened some doors for further research.

REFERENCES

Barsky, R. D., & Lachman, M. E. (1986). Understanding of horizontality in college women: Effects of two training procedures. _International Journal of Behavioral Development_, _9_, 31-43.

Beilin, H., Kagan, J., & Rabinowitz, R. (1966). Effects of verbal and perceptual training on water level representation. _Child Development_, _37_, 317-329.

Berenbaum, S. A., & Resnick, S. (1982). Somatic androgyny and cognitive abilities. _Developmental Psychology_, _18_, 418-423.

Blades, M., & Spencer, C. (1989). Young children's ability to use coordinate references. _Journal of Genetic Psychology_, _150_, 5-18.

Boles, D. B. (1980). X-linkage of spatial ability: A critical review. _Child Development_, _51_, 625-635.

Carlson, N. R. (1980). _Physiology of behavior_ (2nd ed.). Boston: Allyn and Bacon.

Cramer, B. (1971). Sex differences in early childhood. _Child Psychiatry and Human Development_, _1_, 133-151.

De Lisi, R. (1983). Development and individual differences in children's representation of the horizontal coordinate. _Merrill-Palmer Quarterly_, _29_, 179-196.

Diamond, R., Carey, S., & Back, K. J. (1983). Genetic influence on the development of spatial skills during early adolescence. Cognition, 13, 167-185.

Doty, R. L. (1970). The effect of instructions on the middle category of a method of limits investigation of the straight up-and-down position. Journal of Psychology, 76, 9-17.

Geiringer, E. R., & Hyde, J. J. (1976). Sex differences on Piaget's water level task: Spatial ability incognito. Perceptual and Motor Skills, 42, 1323-1328.

Golbeck, S. L. (1986). The role of physical context in Piagetian spatial tasks: Sex differences in spatial knowledge? Journal of Research in Science Teaching, 23, 365-376.

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. Nature, 349, 154-155.

Goodrich, G. A., Damin, P., & Ascione, F. R. (1988, March). Piagetian visual-spatial representation tasks; Performance of kindergarten through fifth graders, college students, and senior citizens. Paper presented at the biennial meeting of the Southwestern Society for Research in Human Development, New Orleans, LA.

Halpern, D. F. (1986). A different answer to the question, "Do sex-related differences in spatial abilities exist?" American Psychologist, 41, 1014-1015.

Harris, L. J., Hanley, C., & Best, C. T. (1978). Conservation of horizontality: Sex differences in sixth-graders and college students. In M. S. Smart & R. C. Smart (Eds.), Adolescence: Development and relationships (pp. 95-107). New York: Macmillan.

Hartston, W. (1987). Drunken goldfish and other irrelevant scientific research. New York: Fawcett Crest.

Humphrey, G. K., Goodale, M. A., & Gurnsey, R. (1991). Orientation discrimination in a visual form agnosic: Evidence from the McCullough effect. Psychological Science, 2, 331-335.

Jamison, W., & Signorella, M. L. (1980). Sex typing and spatial ability: The association between masculinity and success on Piaget's water-level task. Sex Roles, 6, 345-353.

Kalichman, S. C. (1986). Horizontality as a function of sex and academic major. Perceptual and Motor Skills, 63, 903-906.

Kalichman, S. C. (1987, March). Water running-up-hill: A closer look at sex differences in perceiving the invariant horizontality of liquids. Paper presented at the annual meeting of the Southeastern Psychological Association, Atlanta, GA.

Kalichman, S. C. (1988). Individual differences in water-level task performance: A component-skills analysis. Developmental Review, 8, 273-295.

Kalichman, S. C. (1989). Sex roles and sex differences in adult spatial performance. Journal of Genetic Psychology, 150, 93-100.

Kelly, G. N., & Kelly, J. T. (1978). Basic physical concepts: Do college students have them? Journal of College Science Teaching, 8, 29-31.

Kelly, J. T., & Kelly, G. N. (1977). Perception of horizontality by male and female college students. Perceptual and Motor Skills, 44, 724-726.

Kelly, J. T., & Kelly, G. N. (1978). Science activity centers: Basic concepts of the physical world. Science and Children, 16(3), 25-26.

Krasnoff, A. G., Walker, J. T., & Howard, M. (1989). Early sex-linked activities and interests related to spatial abilities. Personality and Individual Differences, 10, 81-85.

Liben, L. S. (1978). Performance on Piagetian spatial tasks as a function of sex, field dependence, and training. Merrill-Palmer Quarterly, 24, 97-110.

Liben, L. S. (1991). Adults' performance on horizontality tasks: Conflicting frames of reference. Developmental Psychology, 27, 285-294.

Liben, L. S., & Golbeck, S. L. (1980). Sex differences in performance on Piagetian spatial tasks: Differences in competence or performance? Child Development, 51, 594-597.

Liben, L. S., & Golbeck, S. L. (1984). Performance on Piagetian horizontality and verticality tasks: Sex-related differences in knowledge of relevant physical phenomena. Developmental Psychology, 20, 595-606.

Liben, L. S., & Golbeck, S. L. (1986). Adults' demonstration of underlying Euclidean concepts in relation to task context. Developmental Psychology, 22, 487-490.

Lunneborg, C. E., & Lunneborg, P. W. (1984). Contribution of sex-differentiated experiences to spatial and mechanical reasoning abilities. Perceptual and Motor Skills, 59, 107-113.

Mackay, C., Brazendale, A., & Wilson, I. (1972). Concepts of horizontal and vertical: A methodological note. Developmental Psychology, 7, 232-237.

Mandler, J. M. (1988). The development of spatial cognition: On topological and Euclidean representation. In J. Stiles-Davis, M. Kritchevsky, & U. Bellugi (Eds.), Spatial cognition: Brain bases and development (pp. 423-432). Hillsdale, NJ: Lawrenece Erlbaum Associates.

Maxwell, J. W., Croake, J. W., & Biddle, A. P. (1975). Sex differences in the comprehension of spatial orientation. Journal of Psychology, 91, 127-131.

McGillicuddy-De Lisi, A. V., De Lisi, R., & Youniss, J. (1978). Representation of the horizontal coordinate with and without liquid. Merrill-Palmer Quarterly, 24, 199-208.

Meehan, A. M. (1984). A meta-analysis of sex differences in formal operational thought. Child Development, 55, 1110-1124.

Meehan, A. M., & Overton, W. F. (1986). Gender differences in expectancies for success and performance on Piagetian spatial tasks. Merrill-Palmer Quarterly, 32, 427-441.

Morris, B. B. (1971). Effects of angle, sex, and cue on adults' perception of the horizontal. Perceptual and Motor Skills, 32, 827-830.

Murphy-Berman, V., Witters, L., & Harding, R. (1986). The effect of bottle shape, bottle position, and subject gender on intermediate-aged hearing-impaired students' performance on the water line task. The Volta Review, 88, 37-46.

Myer, K. A., & Hensley, H. (1984). Cognitive style, gender, and self-report of principle as predictors of adult performance on Piaget's water level task. Journal of Genetic Psychology, 144, 179-183.

Neimark, E. D. (1981). Confounding with cognitive style factors: An artifact explanation for the apparent nonuniversal incidence of formal operations. In I. E. Sigel, D. M. Brodzinsky, & R. M. Golinkoff (Eds.), New directions in Piagetian theory and practice (pp. 177-189). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newcombe, N., & Dubas, J. S. (1987). Individual differences in cognitive ability: Are they related to timing of puberty? In R. M. Lerner & T. T. Foch (Eds.), Biological-psychosocial interactions in early adolescence (pp. 249-302). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newcombe, N., Bandura, M. M., & Taylor, D. G. (1983). Sex differences in spatial ability and spatial activities. Sex Roles, 9, 377-386.

Ohuche, N. M. (1984). Performance on the coordinate reference system: Are gender differences universal? Journal of Cross-Cultural Psychology, 15, 285-296.

Olson, D. M., Eliot, J., & Hardy, R. C. (1988). Relationships between activities and sex-related differences in performance on spatial tests. Perceptual and Motor Skills, 67, 223-232.

Pascual-Leone, J., & Morra, S. (1991). Horizontality of water level: A new-Piagetian developmental review. Advances in Child Development and Behavior, 23, 231-276.

Peskin, J. (1980). Female performance and Inhelder and Piaget's tests of formal operations. Genetic Psychology Monograph, 101, 245-256.

Petersen, A. C. (1976). Physical androgyny and cognitive functioning in adolescence. Developmental Psychology, 12, 524-533.

Petersen, A. C. (1981). Sex differences in performance on spatial tasks: Biopsychosocial influences. In A. Ansara, N. Geschwind, A. Galaburda, M. Albert, & N. Gartrell (Eds.), Sex differences in dyslexia. Towson, MD: The Orton Dyslexia Society.

Petersen, A. C. (1983, August). The development of sex-related differences in achievement. Paper presented at the annual convention of the American Psychological Association, Anaheim, CA. (ERIC Document Reproduction Service No. ED 240 460)

Piaget, J., & Inhelder, B. (1956). The child's conception of space (F. J. Langdon & J. L. Lunzer, Trans.) Atlantic Highlands, NJ: Humanities Press. (Original work published 1948)

Rawson, H., & Miner, M. (1986). The new international dictionary of quotations. New York: E. P. Dutton.

Rebelsky, F. (1964). Adult perception of the horizontal. Perceptual and Motor Skills, 19, 371-374.

Rierdan, J., & Koff, E. (1984). Age at menarche and cognitive functioning. Bulletin of the Psychonomic Society, 22, 174-176.

Sanders, B., & Soares, M. P. (1986). Sexual maturation and spatial ability in college students. Developmental Psychology, 22, 199-203.

Sholl, M. J. (1989). The relation between horizontality and Rod-and-Frame and vestibular nagivational performance. Journal of Experimental Psychology, 15, 110-125.

Signorella, M. L., Jamison, W., & Krupa, M. H. (1989). Predicting spatial performance from gender stereotyping in activity preferences and in self-concept. Developmental Psychology, 25, 89-95.

Signorella, M. L., Krupa, M. H., Jamison, W., & Lyons, N. (1986). A short version of a spatial activity questionnaire. Sex Roles, 14, 475-479.

Smedslund, J. (1963). The effect of observation on children's representation of the spatial orientation of a water surface. Journal of Genetic Psychology, 102, 195-202.

Strauss, E., & Kinsbourne, M. (1981). Does age at menarche affect the ultimate level of verbal and spatial skills? Cortex, 17, 323-326.

Svinicki, J. G., Bundgaard, C. J., Schwensohn, C. H., & Westgor, D. J. (1974). Physical activity and visual field-dependency. Perceptual and Motor Skills, 39, 1237-1238.

Tegano, R. Z. (1982). Sexual differences in the acquisition of Piaget's concepts of horizontality as related to mechanical ability, state anxiety and ACT scores (Doctoral dissertation, University of Sourthern Mississippi, 1982). Dissertation Abstracts International, 42, 3405B.

Thomas, H., & Jamison, W. (1975). On the acquisition of understanding that still water is invariantly horizontal. Merrill-Palmer Quarterly, 21, 31-44.

Thomas, H., & Jamison, W. (1981). A test of the X-linked genetic hypothesis for sex differences on Piaget's water-level task. Developmental Review, 1, 274-283.

Thomas, H., Jamison, W., & Hummel, D. (1973). Observation is insufficient for discovering that the surface of still water is invariantly horizontal. Science, 181, 173-174.

Waber, D. P. (1977). Sex differences in mental abilities, hemispheric lateralization, and rate of physical growth at adolescence. Developmental Psychology, 13, 29-38.

Walker, J. T., & Krasnoff, A. G. (1978). The horizontality principle in young men and women. Perceptual and Motor Skills, 46, 1055-1061.

Willemsen, E., & Reynolds, B. (1973). Sex differences in adults' judgments of the horizontal. Developmental Psychology, 8, 309.

Willemsen, E., Buchholz, A., Budrow, M. S., & Geannacopoulos, N. (1973). Relationships between Witkin's rod-and-frame task and Piaget's water-line task for college women. Perceptual and Motor Skills, 36, 958.

Wittig, M. A., & Allen, M. J. (1984). Measurement of adult performance on Piaget's water horizontality task. Intelligence, 8, 305-313.

APPENDICES

Appendix A: Glossary of Terms

Coordinate axis: A two- or three-dimensional matrix that can be referenced by labelled units along the axes. A coordinate axis system is a cognitive construct which permits an individual to reason within a real or imagined orthogonal matrix.

Euclidean concepts: Used here to refer to the understanding that space can be usefully conceived as a stable, three-dimensional geometric construct. Euclidean concepts are said to be mature when (1) space is perceived in three mutually perpendicular dimensions and (2) one of those dimensions is invariantly perceived as parallel to the pull of gravity.

Euclidean space: A useful but inaccurate conception of space wherein space is composed of three dimensions, each being rectilinear and perpendicular to the other two dimensions.

Horizontal: A line or plane that is perpendicular to the pull of gravity.

Horizontal concepts: The beliefs, predictions, and expectations a person has regarding horizontal phenomena.

Horizontality: (1) The nature and characteristic of being horizontal. (2) Accuracy of horizontal concepts.

Orthogonal: In mathematics this refers to anything which is composed of right angles. It is used here to refer to any line or surface that is parallel to or perpendicular to a given reference direction, usually the pull of gravity or the base of the test paper.

Plumb line: (1) A tool used to determine verticality, composed of a weight (plumb bob) and line. (2) The orientation of a real, imagined, or pictorially rendered resting plumb line.

Plumb line test: A traditional and common way to measure vertical concepts. A subject is shown variously oriented line drawings of structures from which a rope, string, or cord is said to hang. The subject is asked to draw a line indicating the position of the rope. Accuracy is measured by determining the angular deviance of the rendered line from the sides of the page.

Rod and Frame Test: A test used to measure field dependence. In this test subjects are seated upright in a darkened room. They are shown an actual rod and rectangular frame. The frame is presented in several orientations. The subject's task is to orient the rod to an upright position, regardless of the angle of the frame. In addition to measuring field dependence this test taps verticality knowledge.

Spatial concepts: A broad category of cognitive and performance skills that involve solving problems dealing with interrelationships and manipulations of various points, objects, and shapes in space.

Vertical: A line or plane that is parallel to the pull of gravity.

Vertical concepts: The beliefs, predictions, and expectations a person has regarding vertical phenomena.

Verticality: (1) The nature and characteristic of being vertical. (2) Accuracy of vertical concepts.

Verticality/Horizontality measures: The methods whereby a person's understanding of verticality and horizontality can be assessed.

V/H: An abbreviation for verticality and horizontality.

V/H sex difference: Refers to the commonly found difference in mean scores of men and women on measures of verticality and horizontality.

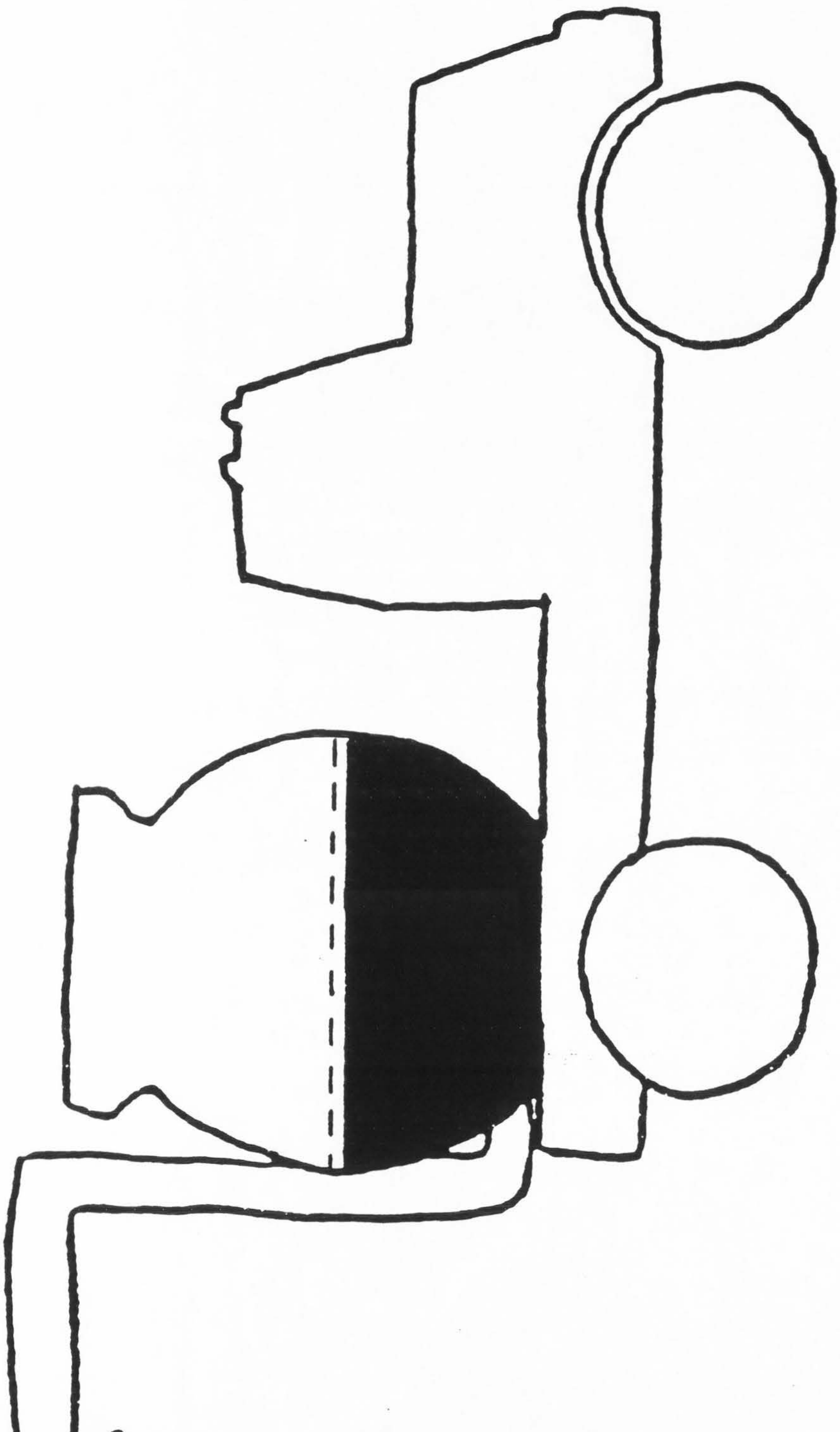
Vertical and horizontal invariance: The fact that horizontality, verticality, and related phenomena are not affected by the orientation of surrounding structures.

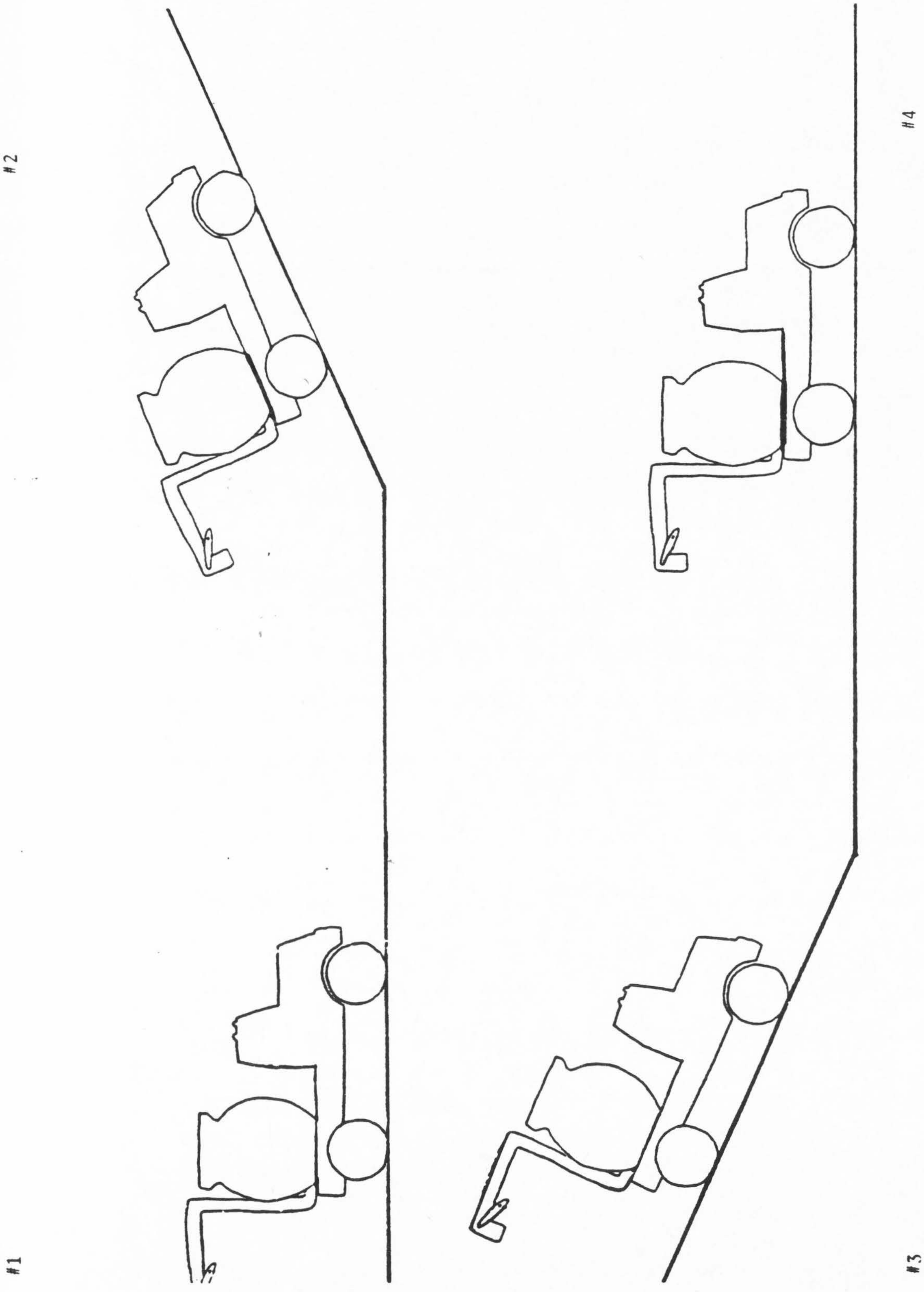
Water level: (1) A tool composed of clear, flexible tubing and colored liquid, used to determine horizontality. (2) The orientation of a real, imagined, or pictorially rendered surface of water or other liquid.

Water-level test: A traditional and common way to measure horizontality. A subject is shown line drawings of variously oriented containers. The subject is asked to draw lines indicating the surface of the water if the container were half full. Accuracy is measured by determining the angular deviance of the rendered line from the top or bottom of the test page.

Appendix B: Root Beer Truck Test

A sample of the root beer truck test is on the following two pages.




#1
#2
#3
#4

Appendix C: Script for Root Beer Truck Test

This is the root beer truck. Every week the root beer truck comes to your town so the people do not run out of root beer. When you look through the clear glass tank you can see that the root beer truck is only half full. At the top of the pop you can see a dotted line. Take your pencil and connect the dotted line, like this. Now go ahead and do that on your paper. Very good. The line you just drew shows the top of the pop.

At the back of the truck is a rope which hangs freely. When you pull the rope it turns on a pump to get the root beer out the back. Take your pencil and connect the dotted line showing the pull rope, like this. And when you are done turn the page.

The second page shows four pictures. This first picture has the root beer truck where it has been parked for 15 minutes. The root beer has stopped sloshing around and the rope has stopped swinging. While it is there take your pencil and draw a line showing the top of the root beer, like this. Remember, it's only half full. And then draw another line for the pull rope at the back. It goes here, like this. Very good.

As you know, in order to get to your town the root beer truck has to go up some hills and then back down. Here we see a picture of the root beer truck on the side of a hill. Because of road construction it has been waiting and waiting for so long that the root beer has stopped sloshing around and the rope has stopped swinging. While the truck is waiting take your pencil and draw a line showing the top of the root beer [pause] and then draw another line at the back showing the pull rope. It goes here. Very good.

Now when the truck gets to the other side of the mountain there's more road construction. And it has been waiting for such a long time that the root beer and the rope have become still. While the truck is waiting draw a line showing the top of the root beer. Remember, it's only half full. And then draw another line for the pull rope, here at the back.

The fourth picture shows the root beer truck parked at your favorite grocery store. It has been there long enough that the root beer has stopped moving around and the rope has stopped swinging. While the truck is waiting take your pencil and draw two more lines, as you know by now, one of them showing the top of the root beer in the tank and another line showing the pull rope at the back.

[Ending #1] When you are done turn the page and complete the demographics questionnaire. If you have any questions you may ask them to the test monitor. Thank you for your participation in this portion of the research.

[Ending #2] You have now completed the root beer truck test. Go ahead and give it to the test monitor.

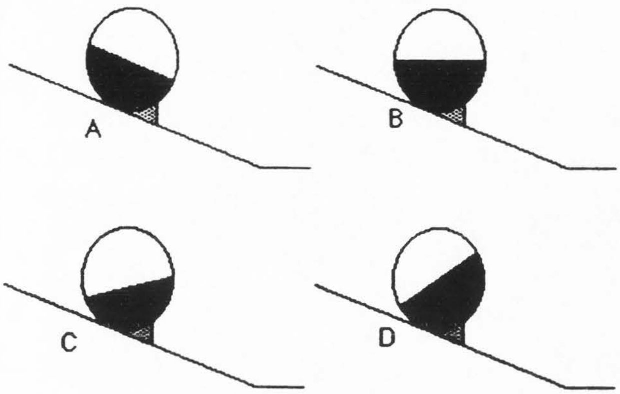
You will be asked to return in two weeks to take the test one more time. Thank you for participating in this research.

Appendix D: Generalization Test

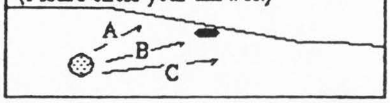
A sample of the generalization test is on the following page.

Last 4 digits of student number:________

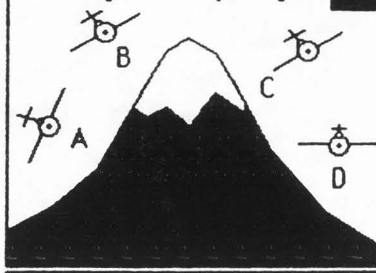
#1 Imagine that someone half-filled a clear glass jar with colored liquid and then set it on the side of a hill. After the liquid stops sloshing around, which of the following would it look like? (Please circle your answer.)



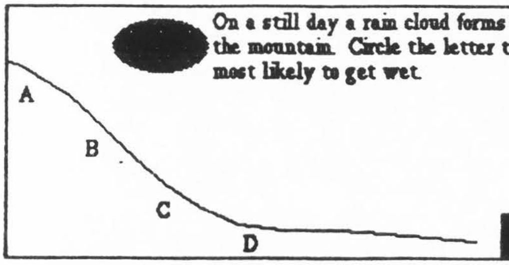
#2 On this sloped putting green which way should the golfer aim to get the ball in the hole? (Please circle your answer.)



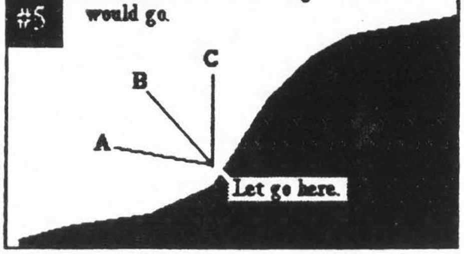
#3 Which airplane will fly straight?



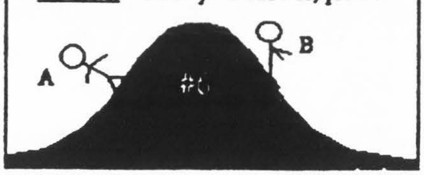
#4 On a still day a rain cloud forms over the mountain. Circle the letter that is most likely to get wet.



#5 If you released a helium balloon from the side of a mountain on a perfectly still day, which way would it float? Circle the letter showing where it would go.

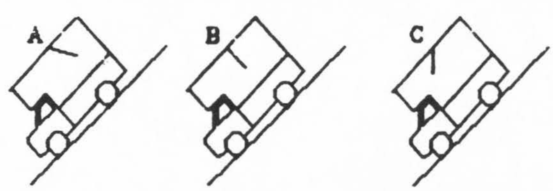


#6 Which stick figure would be most likely to <u>fall down</u>? Circle your answer, please.



#7 Imagine that these glasses had root beer in them and were being held still. Which one must be wrong?

#8 The camper has been parked all night. Inside it there's a string that you can pull to turn on the light. Which picture shows the way the string should look? (Please circle your answer.)



Appendix E: Presentation to Athletes

The following script was read to the athletes to prepare them for taking the V/H test.

Thank you for helping me with my dissertation research. I am studying what and how adults think about a particular type of mental problem. It will take you about 5 minutes to complete this task. By participating you will help me answer some perplexing questions.

Your participation in this research is voluntary. If you do not wish to participate you will not be required to do so. If at any time you wish to stop participating you may.

On the front page of your handout you are asked to give consent to participate. Please read this carefully and, if you agree, sign and date it at the bottom. Please do not put your name on the other pages.

Before I analyze the results I will separate the front page from the rest of the handout. That way your anonymity will be ensured. If you have questions about this research, you may leave a message for me with the receptionist in the Psychology Department or you may get my telephone number from your teacher.

Appendix F: Script for V/H Training

Thank you for participating in my doctoral research. I promise to use your time as effectively as possible. If, at the end, you have any questions, I will answer them for you at that time.

The first thing I want you to do is to put your name on the top line of the index card you were given. Very well, let's begin.

This research deals with two concepts: upright and level. So you will need to know the words that are involved. Vertical is a word which means upright, or straight up and down. It's something you already know, even if you didn't realize it. For example, in order to stand you pretty much have to be vertical. And since you know how to stand, you already know something about the meaning of vertical. People stand vertically. Trees grow vertically. And when there's no wind, rain falls vertically.

Horizontal is a word meaning level, or straight across. [Show graphic showing HORIZON-tal.] It comes from the word horizon [show video of an ocean horizon] and it refers to anything that is perfectly flat and would match an unbroken horizon. Water, when it's still, is always horizontal, or level. Because water stays horizontal it is possible to drink from a glass. And since you know how to drink out of a glass, you already know something about the meaning of the word horizontal.

The words tilted and sloped refer to anything that is not vertical or horizontal, but somewhere in between. [Show picture of a hillside.] Hills and mountains are sloped.

After viewing this tape you will be asked to take the root beer truck test and another related test. In order for you to do well on these two tests, you have to understand what is meant by horizontal and vertical. So pay attention and you will learn how to do it.

Horizontality

Perhaps you, like many other people, have wondered why BYU doesn't have a water skiing team. Well, it's because they can't find any lakes that are tilted enough.

Okay, so it's a dumb joke. But there's a point. Lakes aren't tilted because they're made of water and water likes to be flat. Unless you do something to it, like freeze it, blow on it, or slosh it around, water will always be level. And so all lakes are pretty much flat. You will never see a lake tilted far enough for people to do downhill water skiing. It just won't happen.

If you have a glass, and fill it halfway with liquid, like pop or water, what will it look like? Well that's easy enough--it would look like this. The top of the pop would be perfectly level.

Let's say you get a glass, tilt it, and then fill it halfway with pop. What will it look like? Will the top of the pop tilt with the glass, like this?

We did an experiment to find out. As you can see, when the jar is tipped, the top of the liquid tips too. Does that look right to you? [Pause.] Well it shouldn't, because it's a trick. As you can see, when we show the entire picture, in order to get that shot we had to tilt the camera and pretend to tilt the glass, like this. Now take your 3x5 card and write the following sentence: Still water is always level.

So now you know that a tilted glass would not look like this. Well, some people think that if it's not that way, then maybe the water sloshes to the other direction, so it's higher on the right. Or perhaps it would slope even further than the glass. You may think that the water slopes the same direction as the glass, but only half as far. So if we tilt the bottle straight down the water slopes half way. But does that look right to you? Hopefully it doesn't, because in order to get that shot we had to turn the camera like this. As the bottle tilts one way, the camera goes the other.

Now why does it work that way? Here we will hold the camera straight. Watch what happens. When the bottle tilts down, the red water stays horizontal. When the bottle tilts up, the water is still level. That the nature of water.

Here we have fastened a bottle to a table. When the table is tilted notice what happens to the red water. It stays perfectly level.

*[Enhanced training only.] How can you tell if a thing is horizontal? One easy way is to put a marble on it. If the marble stays in place the thing is horizontal. But if the marble rolls then the thing is slanted and not truly horizontal.

In this case, would the marble roll? No, it would not because it is level. How about this one? Would this marble roll? Yes, it would. That means the surface is sloped. And as you know by now, still water is never sloped. Still water is always horizontal. If you could put a floating marble on it the marble would stay in place. After all, still water is level.

What if you're asked to draw something horizontal; how can you do that? The easiest way is to draw a line so flat that a marble placed on top would not roll. It's that easy.

Verticality

The next point has to do with things that are vertical. A thing is vertical if it goes straight up and down. Many things in life are vertical. Trees grow vertically. Poles are planted vertically. People walk vertically, even when going up and down hills. Weighted lines hang vertically, and the sides of buildings are vertical so that the buildings don't fall down.

Maybe the easiest way to learn about vertical is to think of the epic ballad, "Jack and Jill." As you might recall from you elementary school days, Jack and Jill went up a hill, but eventually came tumbling back down. If you studied carefully you might remember that they are often drawn about like this. If we have them stand still for a moment, notice the angle of their bodies. To make it more obvious let's draw a couple of lines and take away everything else. They are standing on an angle about like this. Now really, is that any way to stand?

To check this out we hired a highly trained, professional stunt stander. Notice how he stands vertically. But, with the help of a rope, he can stand so he's even with the tilt of the hill. As soon as he lets go of the rope he rapidly falls down. Let's try that again in slow motion. Please do not try this at home. In order to stand with the hill he has to hold on to the rope. The very instant he lets go of the rope he begins to fall. That is, he suffers the same fate as Jack and Jill.

So now we know why Jack and Jill fell down the hill. It's because they stood on an angle matching the hill, which of course, caused them to fall.

So ends the tragic tale of Jack and Jill, two who forgot to stand up straight in a crooked world. How might Jack and Jill have stood a better chance? It's easy. You have to stand vertically, even if you're on a hill.

It doesn't matter how things look. Up is always the same direction. And things which are vertical stay in that direction. For example, if it's a still day rain will always fall vertically and balloons will rise vertically, even by a hill. People stand vertically and trees grow vertically, even on a hill.

Toward the ground can mean different directions depending on whether the ground is sloped or flat. The direction down, however, is always the same thing whether the ground is flat or not. Up means the same thing regardless of the slope of the ground. But away from the ground can mean two different things, depending on whether it's sloped or not.

On your 3x5 card write the following sentence: Up is always the same direction.

> What if you're asked to draw a line that is vertical? How can you be sure to do it right? Remember our professional stunt stander. Vertical is unaffected by the slope of hills. Vertical is always straight up and down. Just make you line so it goes straight up and down so that a person standing like that would not fall.

Well by now you know all about vertical and horizontal. As you know, many things in life are horizontal, like lakes, ponds, and good bowling alleys and good pool tables. Many things in life are also vertical, like free hanging lines, trees, even on the sides of hills, rain when there is no wind, buildings so that they don't fall down, and people walking or standing. Remember, many things that are vertical or horizontal are not affected by other things that may be tilted. They will still be vertical or horizontal.

If you are trying to solve a problem about things that are vertical or horizontal remember what you learned in this video.

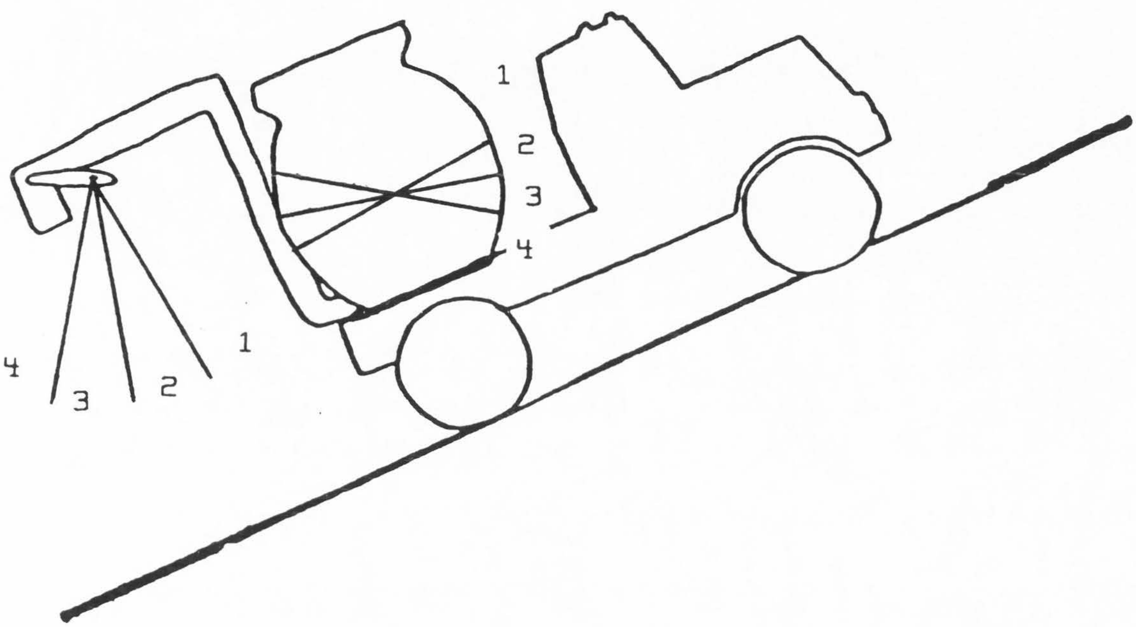
You will now be asked to take the root beer truck test again. Remember what you have learned today and you will do fine. If you have any questions after the test be sure to ask. I will be glad to discuss it with you. Thank you for helping me with my research.

Appendix G: Error Classification Examples

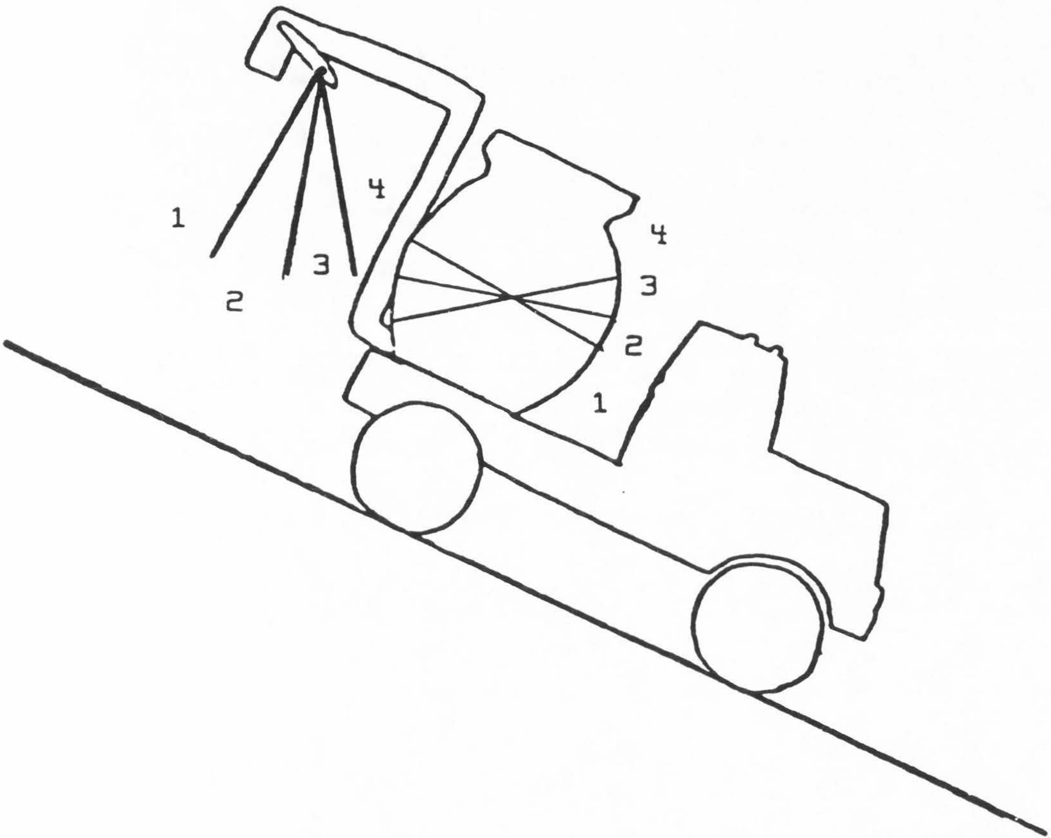
On the following five pages are illustrations and examples of error classification. The first two pages illustrate the V/H error classification system. For the rope, a line falling within area 1, 2, 3, or 4 was a miscorrection, undercorrection, correct answer, or overcorrection, respectively. (The lines in the illustration only indicate boundaries between response categories.)

For the liquid surface the classification system was slightly more complex. Because there was no fixed point that a liquid line must pass through, there was no simple referent for boundaries. Altering the intercept but preserving the slope of rendered lines allowed them to be positioned to pass through the center point of the tank. If the right end of a centered line was in area 1, the response was a miscorrection. If the right end of the line was in area 2 it was an undercorrection. Lines terminating in area 3 or 4 were correct responses or miscorrections.

Following the illustrations of error classification are examples of overcorrections, undercorrections, and miscorrections.

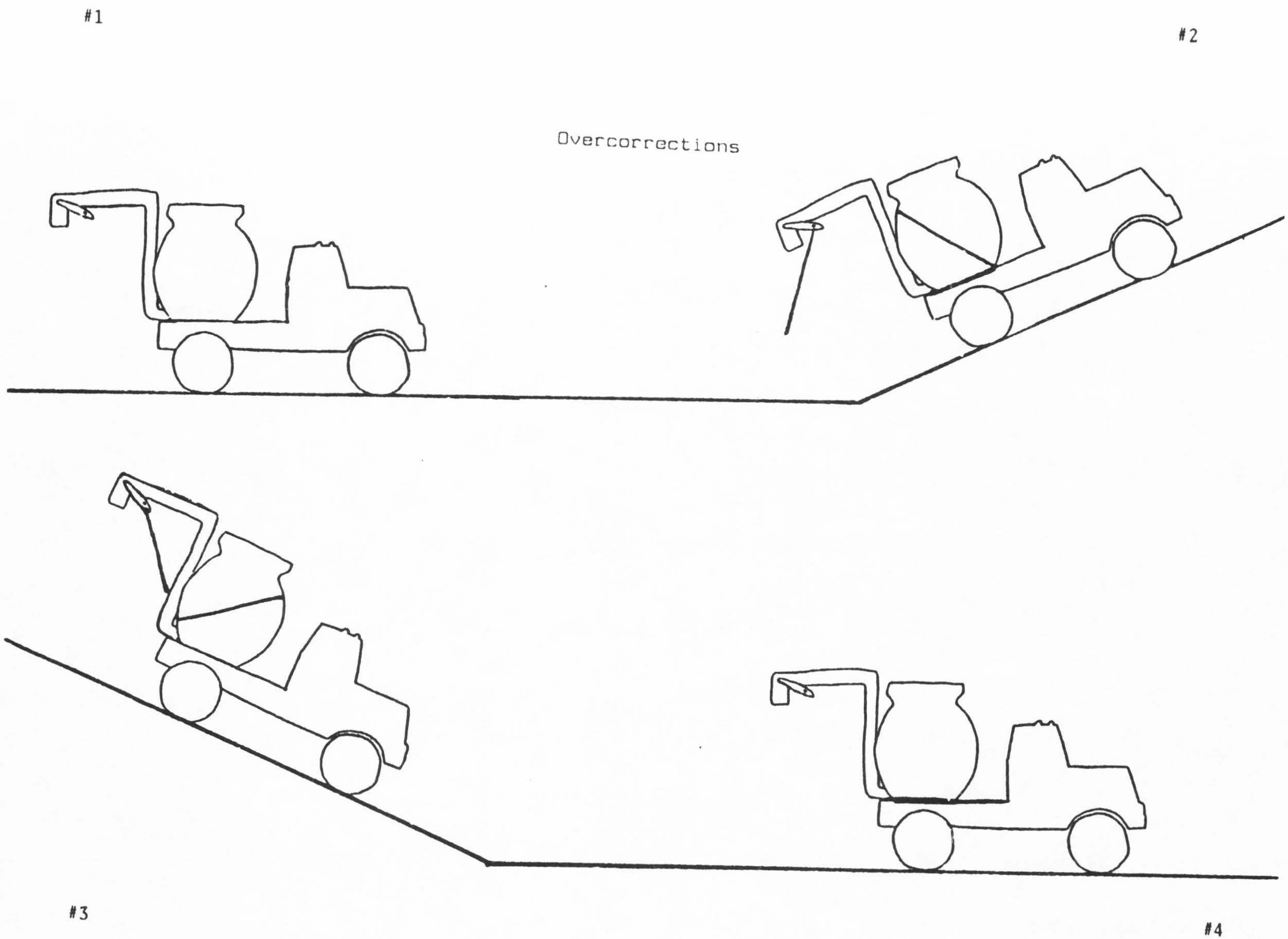


1-Miscorrection
2-Undercorrection
3-Correct
4-Overcorrection



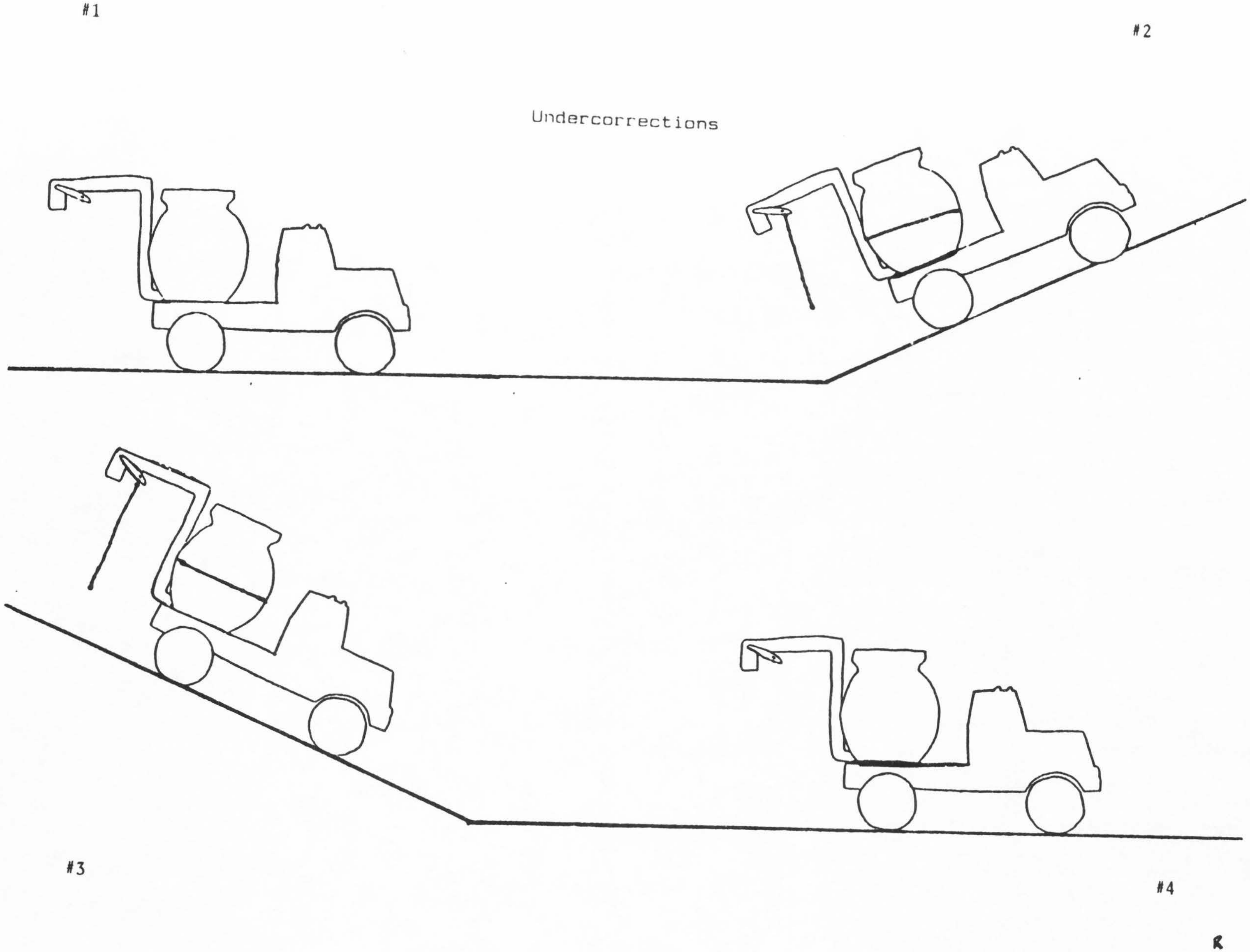
1-Miscorrection
2-Undercorrection
3-Correct
4-Overcorrection

Overcorrections

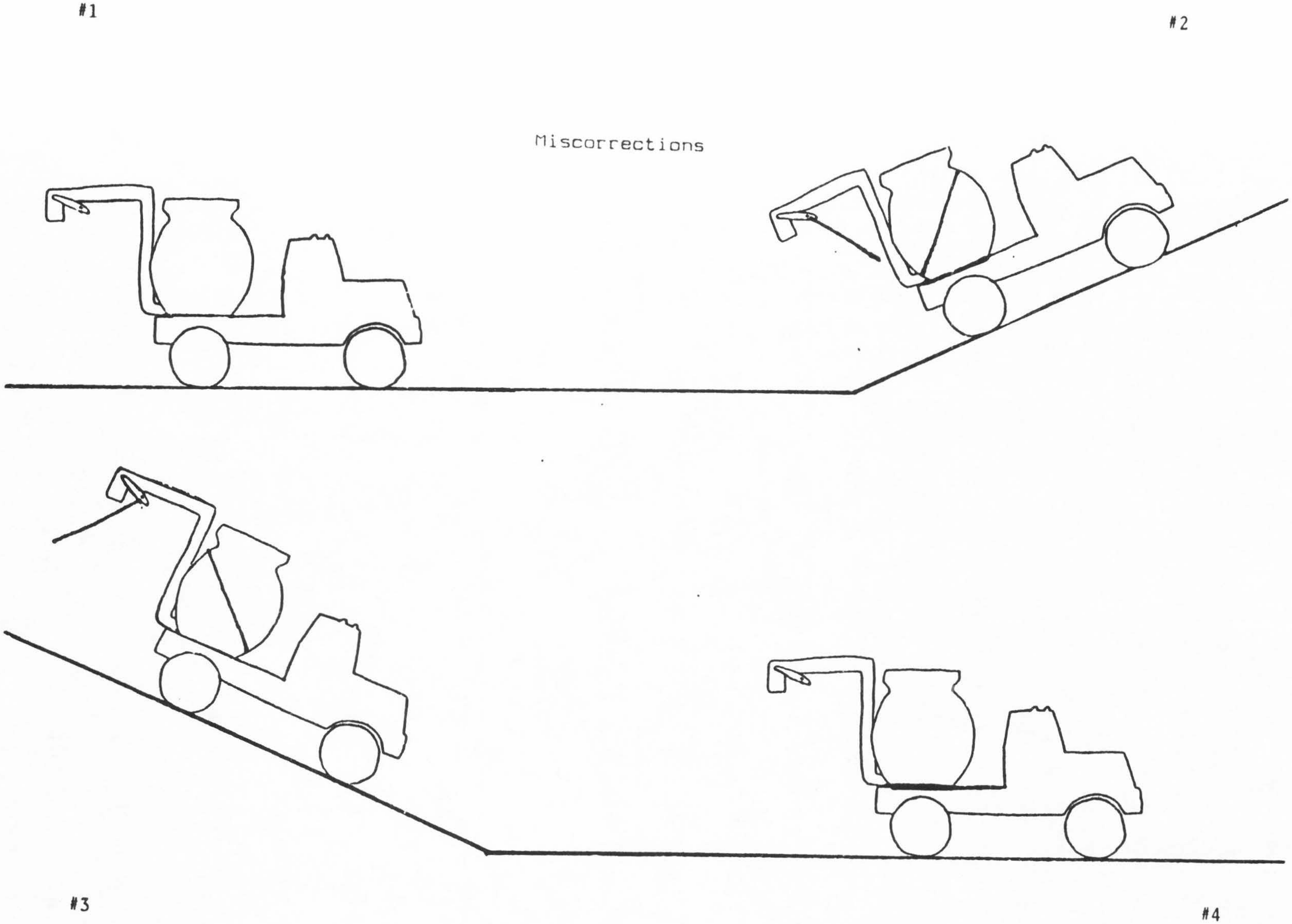


R

Undercorrections



Miscorrections



R

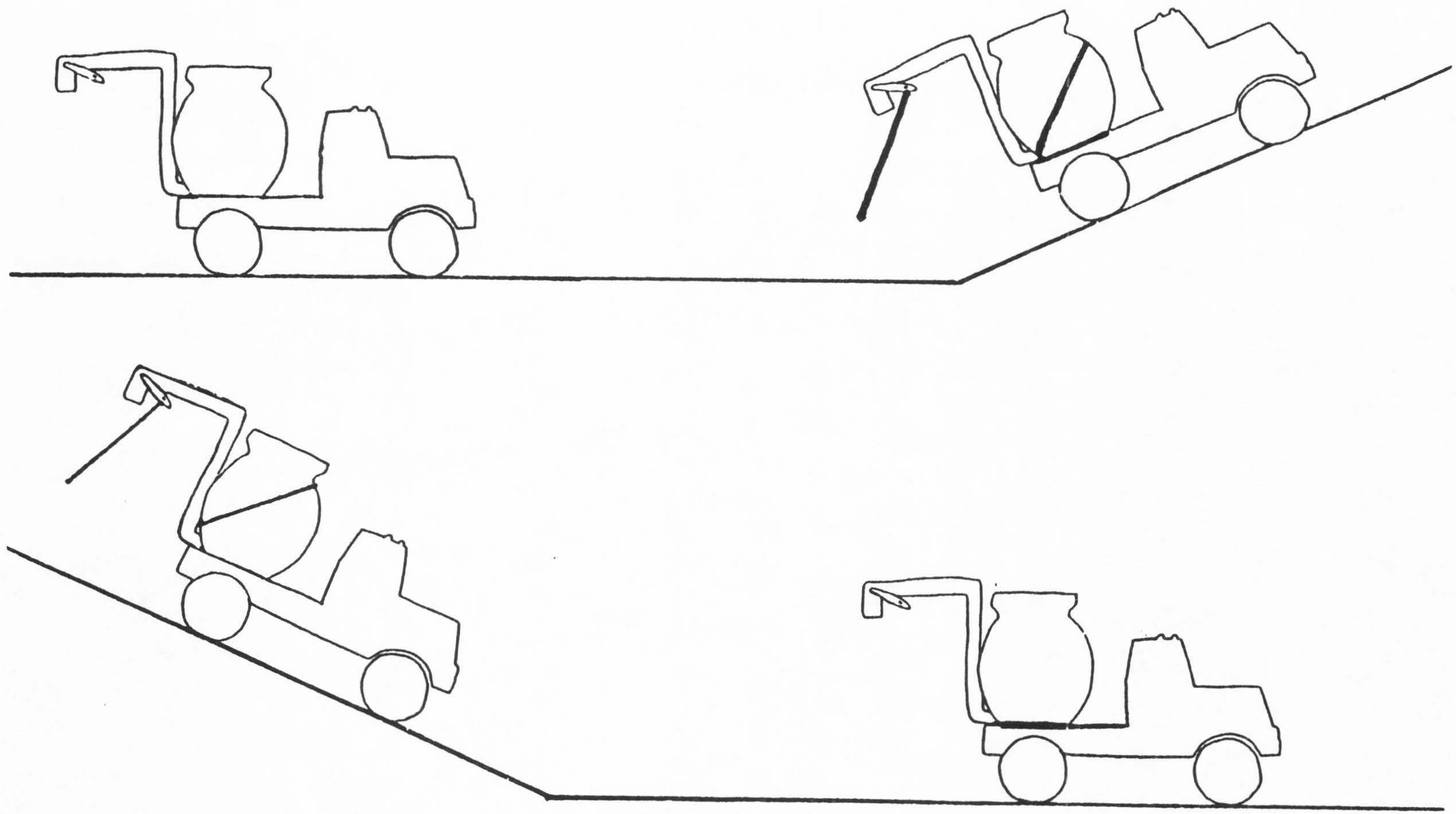
Appendix H: Example of Miscorrection with Overcorrection

On the following page are two examples of the combination of miscorrection and overcorrection errors in a single test stimulus. Notice how the vertical and horizontal referents lose their perpendicularity.

Miscorrections and Overcorrections Combined

#1

#2

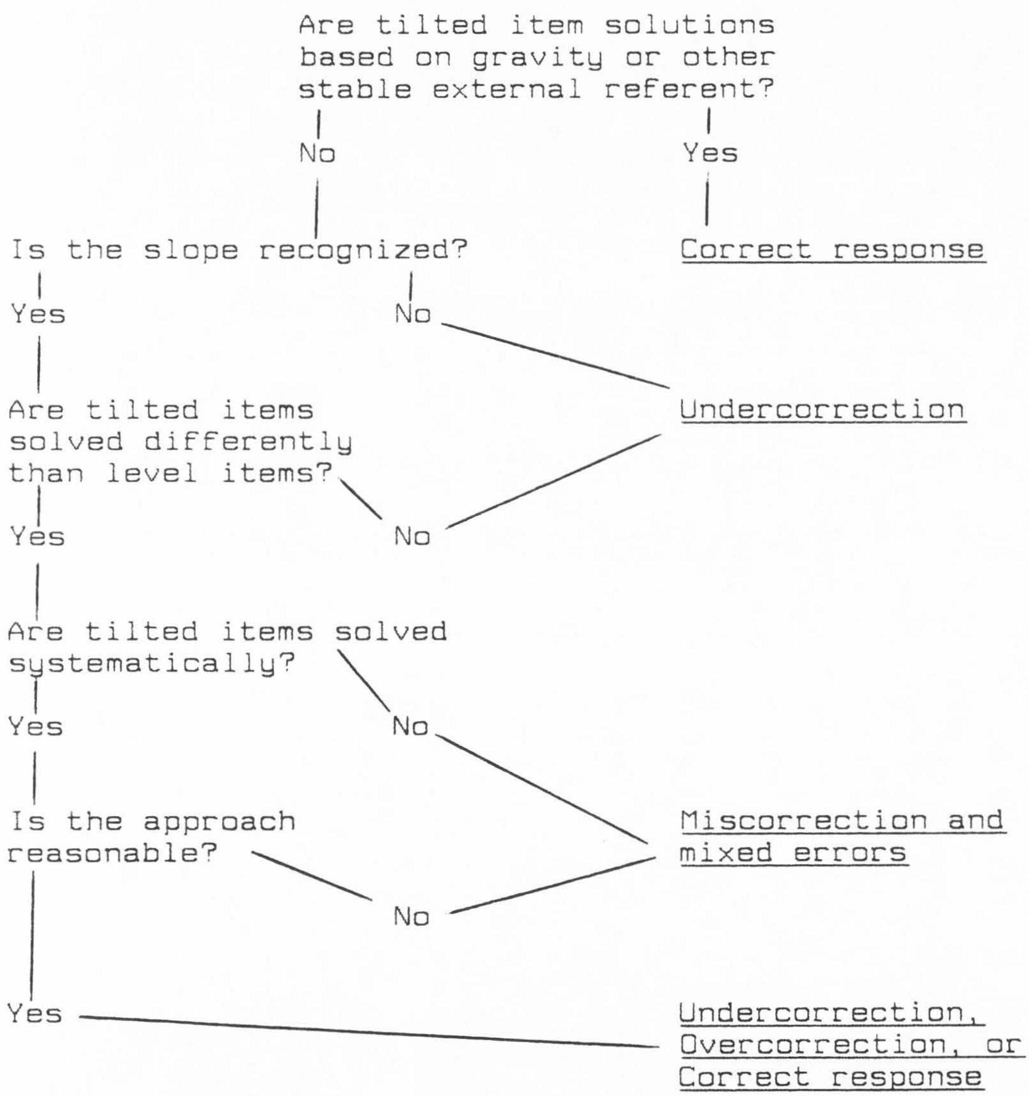


#3

#4

Appendix I: Analysis of Solution Strategies

All subjects performed perfectly on the two test items which were set on level ground. It is therefore extremely unlikely that anyone was performing without a strategy or guiding principle. Because there was no variance in their responses it is impossible to analyze the strategies used. For tilted items, however, there was much variance. What problem solving approaches were used on the tilted test items? Solution strategies can be analyzed according to the following questions.



VITA

GARY A. GOODRICH
UTAH STATE UNIVERSITY
LOGAN, UTAH

EDUCATION

Dissertation: Knowing Which Way is Up: Sex Differences in Understanding Horizontality and Verticality

In 1992 completed requirements for the Doctor of Philosophy degree in Professional/Scientific Psychology at Utah State University in Logan, Utah

Master of Science, Utah State University
June, 1986
Area: Psychology

Bachelor of Science, Brigham Young University
April, 1982
Major: Psychology

WORK EXPERIENCE

Psychology Intern, Salt Lake VA Medical Center.

First Rotation: Outpatient Substance Abuse Clinic, supervised by Dr. Warren Thorley. Duties included testing, intake interviews, individual and marital therapy, group therapy, and psychoeducational classes.

Second Rotation: Inpatient Psychiatric Unit, supervised by Dr. Richard Weaver. Duties included testing, behavior management, short-term therapy, group therapy, psychoeducational classes, and consultation with a multidisciplinary treatment team.

Third Rotation: Neurology Unit, supervised by Dr. Tom Schenkenberg, and Geriatrics, supervised by Dr. Pat Miller. Duties included testing, consultation with patients and staff, brief psychotherapy interventions, and educational presentations to multidisciplinary treatment teams.

Mental Health Extern, Bear River Mental Health.

Supervisor: Dr. Skip Winger.

Date: September 1988 to present.

Duties: Individual, marital, and family therapy, crisis intervention, psychological testing, and any duties requiring Spanish. At BRMH all treatment must be referenced to behavioral goals and therapy emphasis is on time-limited, focused interventions.

Instructor: University Survival and Study Skills.
Supervisor: Dr. Margaret Dyreson
Date: Fall 1988, Fall 1989
Twice taught study skills courses for Academic Services at Utah State University. Focus was on methods of study, knowledge of local resources, and strategies for coping with college life.

Instructor: Introductory Psychology.
For: USU Extension Office
Date: Fall, 1986 through Spring, 1989
Taught for three quarters in Vernal, two quarters for the Ute Indian Reservation Education Office, and one quarter in Moab.

Teaching Assistant: Introductory Psychology.
Supervisors: Dr. Whorton Allen and Linda Barnard
Date: Fall 1986 through Spring 1988
Duties: Teach, tutor, conduct review sessions, and score tests and assignments.

Teaching Assistant: Educational Psychology.
Supervisors: Dr. Margaret Dyreson, David Christian
Date: 10/85 to 5/86, 10/88 to 6/89
Duties: Teach, conduct labs, grade papers and tests.

Research Assistant, Early Intervention Research Institute.
Supervisor: Dr. Karl White.
Date: 11/82 through 9/85, 6/86 through 9/86.
Duties: Development of a meta-analysis coding sheet, meta-analysis coding, monitoring of coding reliability, design of primary research projects, and testing. Also participated in statistical analyses, data quality assurance, and supervision of testing procedures.

PROFESSIONAL PRESENTATIONS

Goodrich, G. A. (1984, June). The Research of Harold Skeels: Contributions to Psychology and Early Education. Paper presented at the annual meeting of the Early Intervention Research Institute, Logan, UT.

Goodrich, G. A., Damin, P. B., and Ascione, F. R. (1988, March). Piagetian visual-spatial representation tasks: Performance of kindergarten through fifth graders, college students, and senior citizens. Paper presented at the biennial meeting of the Southwestern

Society for Research in Human Development, New Orleans, LA.

Damin, P. B., Goodrich, G. A., and Ascione, F. R. (1990, March). Sex role orientation and performance of college-age women and men on Piagetian verticality/horizontality tasks. Paper presented at the biennial meeting of the Society for Research in Human Development, Dallas, TX.

White, K. R., Goodrich, G. A., and Bush, D. (1983, April). The integration of completed research: Procedures and state of the art. Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Snowbird, UT.

WORKSHOPS CONDUCTED

Assertiveness Training, March, 1983, Office Occupations Training Program, USU.

Stress Reduction at USU, presentation to students in the college survival class at USU, Fall 1987.

Stress Management for Teachers, presented to first and second year teachers at Wilson Elementary School, Fall 1988.

Test Anxiety Workshop, 6 separate presentations for USU Academic Services, Winter and Spring quarters, 1989.

INTERESTS AND HOBBIES

An avid fan of music, I play bluegrass, classical, and folk guitar. I like volleyball, softball, cross-country skiing, mountain bikes, frisbees, computers, and video production.