Glenn Koh
Thomas E. von Wiegand
Rebecca Lee Garnett
Nathaniel I. Durlach
durlach@cbgrle.mit.edu
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge MA 02139

Barbara Shinn-Cunningham
Boston University

Use of Virtual Environments for Acquiring Configurational Knowledge about Specific Real-World Spaces:

I. Preliminary Experiment

Abstract

A relatively simple architectural space was modeled and used to compare the effects of spatial training in simulations versus training in the real world. Thirty-five subjects were trained in one of the following conditions: real world (RW), virtual environment (VE), nonimmersive virtual environment (NVE), and model (Mod). The VE condition made use of a head-mounted display to view the simulated environment, while the NVE condition used a desktop monitor. In the Mod condition, the subject viewed and could manipulate a 3-D model of the space, viewed from a desktop display. The training-transfer tasks, performed after brief unstructured exposure to the actual space or to one of the simulations, consisted of estimating the bearing and range to various targets in the real space from various spatially distributed stations, each such pair of estimates constituting a subtask of the overall transfer task. Results obtained from each of the four training conditions proved to be roughly the same. Training in any one of the simulations was comparable to training in the real world. Independent of training condition, there was a strong tendency among subjects to underestimate range. Variability in range errors was dominated by differences among subjects, whereas variability in bearing errors was dominated by differences among subtasks. These results are discussed in the context of plans for future work.

1 Introduction

As evidenced both in this special issue of *Presence* and in a recent past special issue (Vol. 7, No. 2), considerable attention is being given to the use of virtual environment technology for training spatial behavior.

In our research program, we are primarily concerned with the use of VE technology to train individuals' spatial behavior in the real world. We are interested in spatial behavior in virtual worlds only to the extent that it relates to behavior in the real world. More specifically, our focus is on transfer from experiences in the virtual world to behavior in the real world. Furthermore, within this general domain, we are concerned with two distinct but related sets of objectives. The first set focuses on the use of VE technology for familiarizing individuals with specific spaces. The second set focuses on the use of VE technology to improve an individual's spatial behavior in general. In this paper, attention is confined entirely to the first set.

Presence, *Vol. 8, No. 6, December 1999,* 632–656

Within the domain of specific-space familiarization, we intend not only to explore how training that exploits VE technology compares to training conducted in the real world, but also how it compares to training based on simplified and less costly technology. It should also be noted that we do not assume that the best VE trainer is the one that provides the highest fidelity or the greatest degree of realism. Rather, we believe that, as in many other forms of training, simplification, emphasis, and other types of ''distortions'' may enhance training.

This paper is divided into four sections. In Section 2, we have attempted not only to help the reader put our own work in context, but also to provide a brief overview of the wide variety of types of VE / spatial work now being conducted. In Sections 3 and 4, we report on the results of a preliminary experiment that was performed as a first step in our developing program to explore the use of VEs to train spatial behavior in real specific spaces. In Section 5, we summarize our main results and comment on implications and plans for future work.

2 Background

Spatial behavior has been studied extensively for many years in essentially all phyla of the animal kingdom and with respect to all kinds of sensory systems. Even when restricted to humans, the literature on spatial perception, cognition, and behavior—which we shall refer to simply as *spatial behavior*—is enormous.

In this section, we provide an overview of some recent experimental studies that have been concerned both with spatial behavior and with virtual environments. In several of these studies, the focus is the same as in this paper: the use of virtual environments to train spatial behavior in the real world (training transfer). In others, the focus is on improving people's spatial behavior in virtual environments. In still others, virtual environments are regarded primarily as an experimental tool that can be used to improve basic understanding of spatial behavior. Studies in the first category are considered in some detail below. Studies in the second and third categories are considered briefly at the end of this section.

Training and assessment of spatial behavior using virtual environments (primarily concerned with route-following) has been addressed by Witmer, Bailey, and Knerr (1995) and Witmer, Bailey, Knerr, and Parsons (1996). Subjects were trained to follow a specified complex route through a real building and were divided into three groups: a building rehearsal group, a VE rehearsal group, and a symbolic rehearsal group. Half the subjects in each group studied a map before the rehearsal. In the individual assessment phase, subjects were asked about their sense of direction and were given the Building Memory Test (Eckstrom, French, Harmen, and Dermen, 1990). After rehearsal, the VE group completed questionnaires on simulator sickness and the subjective sense of presence. In the study phase, the designated route was studied (route directions, photos of landmarks, maps for the map subgroups) for fifteen minutes. In the rehearsal phase, the building group rehearsed in the real building, the VE group rehearsed in a high-quality VE representation of the building (SGI reality engine, stereoscopic display with a large field of view using a Fakespace Boom2C), and the symbolic group verbally rehearsed the route directions aloud while viewing the landmark photos. In the transfer-test phase, the subjects traversed the route in the real building, stopping to identify destinations along the way. Measures of performance included number of wrong turns, route-traversal times, misidentification of destinations, and distance traveled. Subjects were also required to demonstrate knowledge of the building configuration (even though they had not been trained for this) by estimating distance and direction of various goal sites from various reference locations.

All groups evidenced some route-learning during rehearsal as a function of rehearsal trial number. However, the performance of the VE group was much worse than that of the symbolic group, which in turn was worse than that of the real building group. In the transfer tests, the building group performed best, the VE group second best, and the symbolic group the worst (differences were larger for the wrong-turns measure than for the traversal-time measure). Configuration knowledge was correlated to some degree with gender, building memory test score, and landmark recall; however, it was

not affected by the training method employed. Overall, very little configuration knowledge was gained from the route-learning experience in the VE. (The fact that extensive route knowledge does not necessarily lead to a high level of configurational knowledge has been demonstrated previously in studies of hospital nurses and taxi drivers (Chase, 1983; Moeser, 1988).)

The use of VEs for training firefighters to navigate and operate in specific spaces has been studied by Bliss, Tidwell, and Guest (1997) and by Tate, Sibert and King (1997). In the study by Bliss et al., VE training for route navigation in an unfamiliar building was compared to training by the use of blueprints and to no training. The test task consisted of rescuing a mock baby (a life-sized doll) in the building. Whereas the VE and blueprint groups were instructed to follow the route that had been trained, the no-training group was told to proceed as they would in a real situation when they had no previous experience with the building. All subjects wore goggles sprayed with paint to simulate the degraded visibility that occurs in real fires. The VE system used an SGI RE2 engine, an HMD with a 30 deg. field of view and 100% stereo overlap, a 6-DOF Ascension head tracker, and a mouse to control motion through the VE. The route to be learned was 389 ft. long, and included 16 decision points (with 13 changes of direction). During training, both the VE group and the blueprint group traversed the prescribed route twice under the guidance of verbal instructions from the experimenter. Test performance measures included time to execute the rescue and number of wrong turns. The results of these experiments showed that VE training and blueprint training were roughly equal, and that both these types of training were much superior to no training (for example, the average navigation times for blueprint, VE, and no training were 100, 115, and 177 sec., respectively).

In the study by Tate et al., the environment was the Navy test ship ex-U.S.S. Shadwell, and the subjects were previously trained naval firefighting personnel. In part 1 of the study, the task was purely a navigation task; in part 2, a real firefighting task was included. In both parts, comparisons were made between traditional mission review and rehearsal, and traditional mission review and rehearsal plus VE rehearsal. The VE system made use of an SGI RE2, a VR4 HMD, a Polhemus tracking device, joystick motion control, and a ''glove avatar'' which, together with the joystick, provided both a ''fly where you point'' metaphor for movement control and a means for opening and closing doors. The VE rehearsal involved a ''magic carpet ride'' through the space, a user-controlled trip through the space, and a user-controlled trip in which visibility was reduced by the introduction of simulated fire and smoke. The results of the study showed that supplementing the traditional review and rehearsal procedures by the VE rehearsal improved performance. In part 1, the VE group took 1 min. 54 sec. to complete the task, whereas the non-VE group took 2 min. 38 sec.; in part 2, the VE group took 9 min. 26 sec. to extinguish the fire, whereas the non-VE group took 11 min. 43 sec. In both parts, the VE group took many fewer wrong turns than the non-VE group.

Darken and Banker (1998) explored the use of VEs for terrain familiarization in natural environments, using experienced male map readers with varying amounts of experience in the sport of orienteering. The test task was to locate nine control points in a specified order, all of which were located in a 1200 m $\times$ 700 m portion of the former Fort Ord in California. Three training methods were employed: map only, VE plus map, and real world plus map. The map was an orienteering map that showed the environment in great detail. The VE had high visual fidelity, used a desktop PC with a mouse and keyboard, and a very low frame rate (3 frames / sec). The views included a top-down map view with a you-are-here indicator, as well as an egocentric view. All subjects had one hour to study the terrain and prepare for the test. At the start of the test, they had access to a map and compass for one minute to allow for initial orientation, and they were permitted at any time to ask for the map or compass for a 30 sec. interval. A differential geographic positioning system (15 m accuracy) was used to collect positional data on the subjects every two seconds. Although interpreting the data with confidence was relatively difficult, the authors concluded that ability level was more important than training method in determining performance, intermediate ability-level users benefited most from the VE training, and the ability to compress travel time in VEs—and thus to experience the

traversal of greater areas in a given time interval—may provide an important advantage to VE training over real-world training.

In a related study, Goerger, Darken, Boyd, Gagnon et al. (1998) considered an architectural rather than natural environment, as well as effects of exposure duration. Whereas subjects in the map group studied floor plans of a seven-story building for thirty minutes, subjects in the map-plus-VE group were simultaneously exposed to a VE representation of the building using a three-screen, rear-projection system with a 145 deg. FOV and a movement control system that made use of verbal commands. Inasmuch as the study period for the latter group was the same as for the former (thirty minutes), time spent on the VE subtracted from the time spent on the floor plans for this group. Thus, although the map-plus-VE subjects had the additional potential benefit of the VE, they had to pay for this potential benefit by decreasing the time used for studying the plans.

Transfer tests in the real seven-story building involved navigating from an entry point through a sequence of four target locations. In addition, subjects were required to point to various target locations (egocentric estimates) and to construct and navigate a path back to the entry point from the last target location. Finally, after completing these tasks, subjects were required to estimate relative distances and bearings among the targets by placing numbered magnets on a metal whiteboard (exocentric estimates).

The results of this study showed that the map group made fewer route-following errors than the map-plus-VE group, that errors in estimation of direction were roughly the same for the two groups (errors on the egocentric task were roughly double that on the exocentric task for both groups), and that the map group was significantly more accurate than the other group in the distance-estimation tasks. The extent to which the performance of the map-plus-VE group relative to that of the map group would have improved if the study time had been much longer than thirty minutes (that is, long enough for the benefits of spending additional time on the floor plan to saturate) was not determined.

In a further related study, Goerger (1998) studied the differential effects of three training conditions on the acquisition of spatial knowledge and navigation performance in the natural-terrain orienteering course at the former Fort Ord. The map group was trained by studying the detailed orienteering map; the real-world group was trained by both studying this map and exploring the actual terrain; and the VE group was trained by both studying this map and making use of a VE representation of the terrain. The map used by all groups was specially developed using a digital aerial photo and ground reconnaissance and was unusually detailed, making use of orienteering symbols and a highly magnified scale (1:5,000 compared to traditional military maps of scale 1:25,000 and to competition orienteering maps of scale 1:15,000). The VE system was high-end, making use of an SGI RE-2 workstation, a three-screen rear projection display system with an FOV of roughly 100 deg., and a 6-DOF joystick for control of movement and orientation. Included in the viewing options was a top-down view, as well as teleportation to a variety of locations within the virtual terrain. Fifteen male subjects (all but one in military service) were given various tests of spatial abilities and then distributed evenly among the three training groups according to results obtained with the Guilford-Zimmerman Spatial Orientation Aptitude Survey (GZ).

In the training phase of the experiment, the subjects were given one hour to study a variety of materials (such as the map, the list of tasks to be performed, and a course clue sheet), and to plan a route to navigate through nine specified control points in a specified order. Whereas the real-world group was given the opportunity to explore the real terrain, the VE group was given the opportunity to explore the virtual terrain. However, as in the study cited above, the total amount of training time was the same for all groups.

In the testing phase of the experiment, the subject had to perform nine planned tasks (locating the nine control points in the proper order) as well as three unplanned tasks of the same type. Evaluation of route-following performance included measurement of the number of deviations and the distance deviated from the planned routes. Subject behavior was recorded by use of a differential global positioning system and a helmet-mounted camera, as well as by direct observation by human ob-

servers. Determination of survey knowledge was based on directional pointing tasks, unplanned route-selection tasks, unplanned navigation tasks, and exocentric spatial tasks using the magnet-whiteboard technique used in the Goerger et al. (1998) study.

Overall, the results failed to show any substantial benefits of using a portion of the training time to explore the real world or the virtual environment; participants in the map group tended to do at least as well as participants in either the real-world group or the VE group. Whereas the real-world group tended to outperform the VE group with respect to route knowledge, the VE group evidenced slightly better performance than the real-world group with respect to survey knowledge. It was also found that the results of the tests of spatial abilities applied were positively correlated with the results obtained in the natural terrain spatial tasks considered.

In a study by Waller, Hunt, and Knapp (1998), the effects of different training conditions on spatial knowledge of a 14 ft. × 18 ft. maze was explored. Test tasks in this study included navigating the maze while blindfolded and taking a true-false test in which subjects were required to determine whether a given map correctly represented a portion of the maze. Six training conditions were considered: (1) blind (no training), (2) real-world training, (3) map training, (4) VE desktop training, (5) VE immersive training (2 min. exposure), and (6) VE long immersive training (15 min exposure).

The results of this study for the blindfolded navigation task (measured in terms of time to completion of task) suggested that all the forms of training tested were useful, and that the real-world and VE-long training were somewhat better than the map, desktop, or VE-short training. The results for the true-false test showed best performance for the map group and worst for the VE-short group. Although there was a slight gender effect in the true-false test (men 69% correct, women 64% correct), and a substantial gender effect in the maze performance for the VE training conditions, there was no gender effect in the maze performance for the real-world training condition. It is also worth noting that the Guilford Zimmerman Test of Spatial Orientation was not predictive of performance for most of the subjects.

In addition to the above studies, there have been many studies of spatial behavior that involved the use of VEs but were not explicitly concerned with training transfer to the real world. Brief comments on many of these studies are provided in the following paragraphs.

Henry (1992), concerned with spatial perception in architectural environments, found judgments of room size to be smaller and judgments of angle to be more variable in a simulated environment than in the real environment (although systematic biases in angular judgments were found to occur in the real environment as well as the simulated one).

Peruch, Vercher, and Gauthier (1995), in a study comparing spatial knowledge in a VE acquired through active free exploration with that acquired through passive viewing, found (consistent with much classical work) that active exploration was superior.

May, Peruch, and Savoyant (1995) showed that map misalignment (deviations from the natural correspondence between ‘‘up’’ and ‘‘forward’’) reduced speed and accuracy of navigation in a VE.

Satalich (1995) compared a number of methods for exploring a VE (self-exploration, active guided, passive guided) to a control condition in which subjects only had access to a map of the space. Results on orientation tasks, distance-estimation tasks, and wayfaring tasks in the VE generally showed that performance with the control condition was better than or equal to performance under any of the conditions involving experience in the VE.

Aginsky, Harris, Rensink, and Beusmans (1996) studied route-learning in a driving simulator. They found the environmental information in the vicinity of choice points is more likely to be retained than information at other points and that two strategies emerged: a ‘‘visual strategy’’ that relies primarily on visual recognition of actual intersections along the route and a ‘‘spatial strategy’’ that relies on a mental map that incorporates the environment’s spatial structure.

Darken and Sibert (1993, 1996a, 1996b) performed a variety of studies directed towards improved navigation in VEs, particularly large-scale ones. They developed a navigation tool set and examined the use of various envi-

ronmental cues and navigational aids in a complex nautical search task, using performance measures that included search time, percentage area searched, and errors made in map production. In general, they found that principles previously developed in connection with navigation in real-world spaces could be usefully applied to the problem of navigating in VEs.

Stoakley, Conway, and Pausch (1995) implemented and explored the use of a virtual hand-held miniature model of the VE (referred to as a ''world in miniature,'' or WIM) that provides the user with an exocentric view of the VE that can be changed at will by simply rotating or zooming in on the model. In many respects, a WIM can be regarded as an extension and elaboration of a map that effectively exploits modern VE technology. Among the many ways in which a WIM can be exploited in a VE (a large selection of which are discussed by the authors), some are clearly useful for improving spatial behavior in the VE.

Tlauka and Wilson (1996), interested in differences in spatial knowledge achieved by means of maps and by means of navigation through the space mapped, found that bearing-judgment errors by the map-group subjects were smaller than those by the navigation-group subjects (in contrast the results of Thorndyke and Hayes-Roth (1982) who found directional judgments to be less accurate for a map group than a navigation group in a real-world environment), but that there was no significant effect of alignment in either group.

Pausch, Proffitt, and Williams (1997) examined the relative effectiveness of immersive HMD displays and desktop displays when searching for targets in a heavily camouflaged room. It was found that task completion time (time to find target) was roughly the same for the two types of display on trials in which the target was present, but that it was smaller for the HMD display on trials when no target was present (time to search the entire room and conclude that no target was present).

Colle and Reid (1998), concerned with the acquisition of configuration knowledge from exploration and navigation in a VE, found that recall of direction information was more accurate when the situation to be recalled was one in which the object to be pointed at by the subject was in the same virtual room as the subject rather than in a separate room (the so-called ''room effect''). It was also found that pointing responses based on the analysis of maps that the subjects had drawn following their experiences in the VE were more accurate than those based simply on recall of the VE experience itself.

Witmer and Kline (1998), in a study concerned with the perception of distance in virtual environments, examined distance estimates achieved from static visual cues and from cues obtained by traversing the distance to be estimated. In the static case, estimates of the virtual target distance were only 47% of the ''correct'' virtual distance compared to a figure of 72% in comparable real-world tests. However, the scatter of the data in the VE tests was smaller than in the real-world tests. In terms of resolving power (as opposed to response bias), the VE results appear to have been at least as good as the real-world results. In the distance-traversed experiments (with traversal achieved by means of a joystick, a treadmill, or teleportation), the underestimation of distance was less severe; however, surprisingly, transversal by means of the treadmill did not produce better results than transversal by joystick or by teleportation. It was also found, as one would expect, that an auditory cue that reported distance traveled (one beep every ten feet of movement), drastically reduced both the bias and the scatter evident in the data.

Ruddle, Payne, and Jones (1997, 1998, 1999), concerned with navigation in large and complex virtual buildings, examined a variety of effects, including the effects of navigational experience and orientation aids. In addition, they compared the performance obtained when using an HMD to that obtained when using a desktop monitor.

In the 1997 study, an experiment was performed using a desktop VE system that paralleled in considerable detail the experiment conducted previously in the real world by Thorndyke and Hayes-Roth (1982). Among other things, Ruddle et al. found that route-finding ability improved substantially with navigation practice in the VE, that such experience produced more-accurate estimates of VE route distance than VE Euclidean distance,

that estimates of direction and distance based on navigation experience were slightly less accurate than the equivalent results in the Thorndyke and Hayes-Roth experiment, that subjects with map experience rather than navigation experience made more-accurate judgments of Euclidean distance in the VE, and that the accuracy of the map group estimates in the VE were roughly equivalent for both distance and direction to those of the map group in the Thorndyke and Hayes-Roth study. The navigation group in the VE study, although it showed great variability in the distance estimates, showed no overall tendency to underestimate or overestimate distance; however, these subjects were provided with specific scale information (certain distances were specified in feet or meters). Two further experiments in the 1997 study focused on the benefits provided by different types of landmarks. Based on all three experiments, the authors concluded that alternative navigational aids, beyond presentation of landmark cues, need to be developed for successfully navigating in VEs. Included as candidates for such a purpose were virtual suns, compasses, VE maps, and WIMS (Stoakley et al., 1995).

In the first experiment in the Ruddle et al. 1998 study, a within-subjects design was used to study errors in orientation estimates for simple paths containing one, two, or three turns. Results showed an increase in mean error from about 20 deg. to 33 deg. as the number of turns increased, independent of whether the VE desktop system used employed a 45 deg. or 90 deg. field of view (compared to a mean error of 29 deg. in the 1997 study). In a second experiment, performed in a large-scale VE using the 90 deg. FOV, it was found that route-finding ability and survey knowledge (VE orientation and VE Euclidean distance) improved with experience in a given VE, as well as transferring to some extent from a first to a second (distinct but somewhat similar) VE. However, providing a compass to the subjects had no effect on the various performance measurements, although the availability of the compass did appear to influence certain aspects of behavior during the tests and also the overall comfort of the subjects. In discussing these results, the authors suggest that the results of Presson and Montello (1994) and Rieser (1989) concerning differences in the maintenance of directional information after real or imagined rotations might imply perhaps that global orientation can be better maintained in immersive VEs than in desktop VEs. However, they point out that, in one of their own previous studies (Ruddle, Randall, Payne, & Jones, 1996), route-following performance and accuracy of direction estimates were roughly the same for the two kinds of systems.

In the 1999 study, performance in navigating complex building VEs with an HMD was compared to performance using a desktop monitor. With the HMD (VR4, 247 $\times$ 230 resolution, Polhemus head tracker, 50 deg. FOV), direction and speed of movement was controlled by means of a hand-held button box. With the desktop display (21 in., 1280 $\times$ 1024 resolution, 40 deg. FOV), viewing direction was controlled by means of a mouse and movement direction and speed by use of a keyboard. With both displays, viewing direction and movement direction were decoupled. No stereographics were employed in either display. Each of twelve subjects first navigated around one virtual building four times using one display (HMD or desktop) and then navigated around a second virtual building four times using the other display (with appropriate counterbalancing of display order, building order, and display / building combination). The task required of the subject in each of the four runs assumed greater familiarity with the building as the run number increased from 1 to 4, and included estimation of directions and straight-line distances during run 4. The results of this study showed that the HMD led to slightly more efficient route-following behavior in terms of travel time, but not in terms of travel distance. Increased travel time with the desktop display appeared to be related to an increased tendency to stop before altering direction. With respect to acquisition of survey knowledge, the HMD was found to produce more-accurate estimates of relative straight-line distances. (There was no consistent tendency to underestimate or overestimate these distances with either display.) However, the mean orientation errors observed (which, in general, were quite large, ranging from approximately 40 deg. to 65 deg.) were insignificantly smaller for the HMD case.

Chance, Gaunet, Beall, and Loomis (1998) examined

how directional estimates in virtual mazes are affected by the mode of locomotion used to move through the maze. In the walk mode, subjects walked through the maze wearing an HMD with tracking information being used to update the visual imagery. In the visual-turn mode, subjects moved through the virtual maze using a joystick to control turns. In the real-turn mode (regarded as intermediate between the first two modes), subjects physically turned in place to steer while translating through the maze. Although the errors in the absolute directional estimates (from terminal position to various target positions) tended to be very large (often exceeding 60 deg. in tests for which totally random behavior would have led to a mean absolute error of 90 deg.), there was some indication that proprioceptive information was useful in making the directional estimates (for example, the errors for the walk mode tended to be somewhat smaller than the errors for the visual-turn mode). The walk mode was also less likely to cause simulator sickness in the tests.

Loomis, Golledge, and Klatzky (1998), in a study directed towards the development of a navigation aid for the visually impaired, examined guidance performance as a function of four display modes, one using spatialized sound from a virtual acoustic display and three involving verbal commands issued by a synthetic speech display. Using time-to-complete-course and distance-traveled as performance measures, they found that performance was best with the virtual acoustic display, and that providing verbal guidance information without appropriate heading information caused a significant degradation in performance.

In general, the number of studies concerned with spatial behavior that involve the use of VEs is increasing very rapidly. However, as indicated above, most of these studies focus on improving navigation in VEs and / or on using VEs as a research tool for studying spatial behavior. Relatively few are focused on the use of VEs to train spatial behavior in the real world (either in a specific space or in general) or, equivalently, on issues related to training transfer to the real world. The extent to which the development of methods or aids that improve spatial behavior in VEs (but are not available to the subject in the real world) will lead to improved spatial behavior in the real world, is, of course, an open question and is likely to depend on the specific method or aid in question. In addition, most of the previous work on transfer to the real world is focused on route knowledge, not configuration knowledge. And little attention has been given to the use of 3-D maps or WIMs in the VE training.

3 The Experiment

3.1 The Real Space

The space chosen for our preliminary study consisted of a portion of the seventh floor of Building 36 at MIT. This space, the layout of which is shown in Figure 1, was chosen for convenience, as it is easily accessible and close to the Virtual Environment facilities in our laboratory. Although the space is relatively simple (compared, for example, to the complex spaces employed by Ruddle et al.), we believed it was adequate for our preliminary work.

3.2 The Test Task

The test task employed was chosen to probe the subjects' acquisition of configurational knowledge. In general, we regard configurational knowledge as more important than route knowledge for two reasons. First, there exist tasks for which performance depends more on configurational knowledge than on route knowledge (such as designing a ventilation system for a building or planning to demolish a building). Second, whereas configurational knowledge may imply route knowledge (including knowledge of secondary routes when primary routes are blocked), route knowledge does not generally imply configurational knowledge.[1]

For all training conditions considered (see Section 3.3), the subject, located at a given reference position (''station'') within the real space, was required to esti-

1. One could argue that configurational knowledge does not necessarily imply familiarity with all landmarks relevant to route knowledge, or that route knowledge, if sufficiently complete and sufficiently metric, does imply extensive configurational knowledge. However, we believe that, at least to a first-order approximation, our statement is correct.

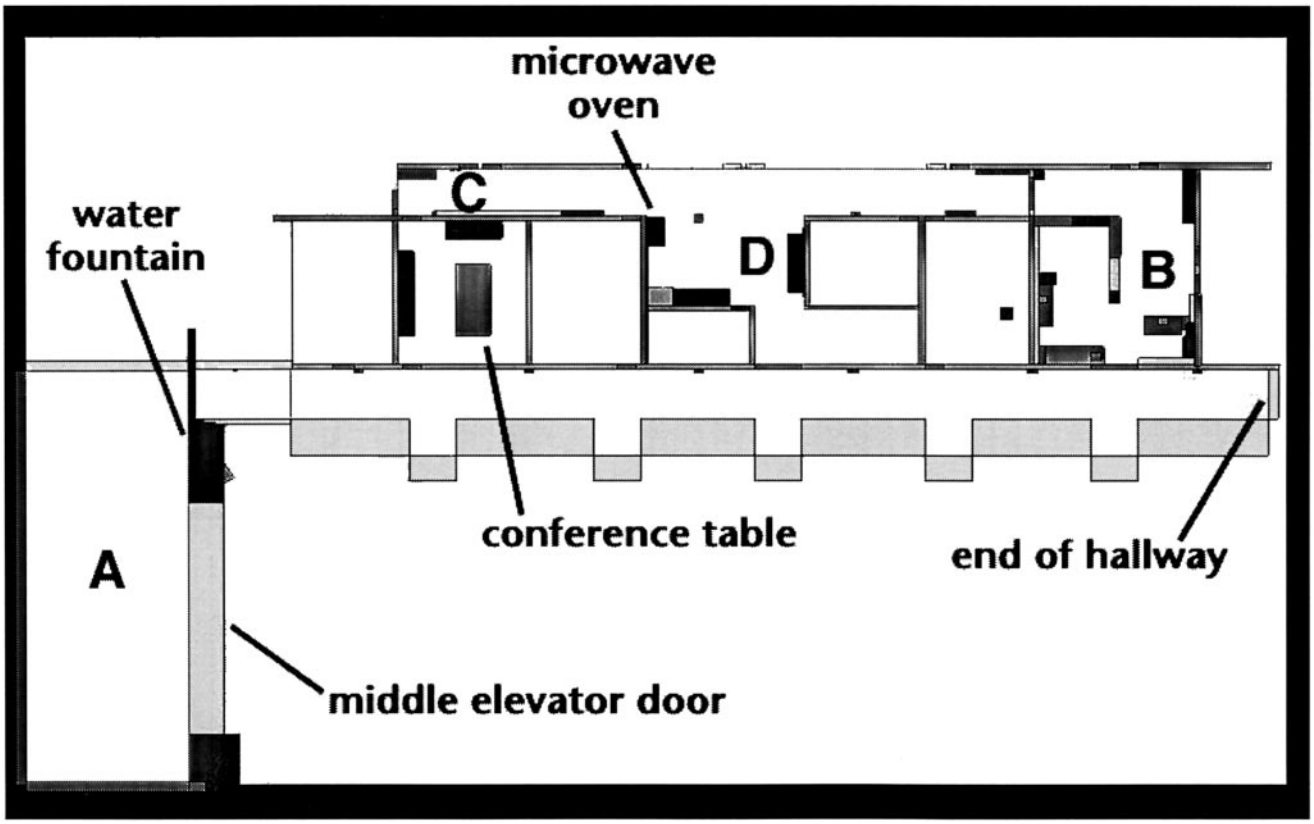

Figure 1. *Overhead view of space used in the experiment with indications of both landmark and reference locations (A, B, C, D).*

mate the location of a given landmark within the real space by reporting the bearing (θ) and the range (*R*) of the landmark relative to the location of the station. Thus, the type of knowledge being probed in these tests was Euclidean-metric configurational, rather than, for example, topological or city-block-metric configurational.

The stations and landmarks used are shown in Figure 1. Of the twenty possible pairs of stations and landmarks (four reference locations times five landmarks), only the thirteen shown in Table 1 were used in the experiment. In all these cases, the landmarks were obscured from view by walls.[2] The various (station, landmark) pairs will be referred to as *subtasks* and identified by number, consistent with the numerical assignments shown in Table 1.

2. Responses were also collected for the case (A, Water Fountain) as a benchmark, the only case in which the landmark was visible from the station. These data are discussed at the end of Section 4.

Table 1. *Subtasks*

	Water fountain	Micro-wave oven	Con-ference table	End of hallway	Elevator door
A		1	2	3	
B		4		5	6
C	7	8		9	10
D			11	12	13

During testing, subjects were brought to the stations in the order in which they are labeled (A, B, C, D) and were transported from one station to the next in a wheelchair while blindfolded. The blindfold was then removed during a subject's attempt to estimate the location of the landmarks. Directional estimation was accomplished by use of a pointing rod and protractor

mounted on a tripod. Feedback was not provided to subjects about errors made in estimates of bearing or range. The total time consumed by each subject to perform the fourteen subtasks was less than one hour. The order in which the subtasks were performed is that shown in Table 1.

3.3 The Training Conditions

Four training conditions were employed, each with a different set of subjects:

1. RW: Real World
2. VE: Immersive VE
3. NVE: Non-Immersive VE
4. Mod: Model

For all these training conditions, the subjects were informed about the task they would be asked to perform after the training was completed, although they were not informed about the specific locations to be used as references or the specific objects to be used as landmarks. For all conditions, training included ten minutes of free exploration under the specified training condition. For the RW condition, this was the only training provided. In the other three cases, an opportunity to become familiar with the technology was provided prior to the training period in which the synthetic version of the seventh-floor space was explored. Subjects in these groups were familiarized with the equipment by permitting them to explore a vastly simplified VE unrelated to the test space.

In the VE condition, subjects used an HMD for visualization and a joystick for navigation. The NVE condition used the same equipment and procedures, except that a 21 in. monitor was used to view the simulation, rather than the HMD. For both conditions, the joystick was configured so that a forward push moved the subject forward in the scene, a backward push moved the subject backward, and left and right pushes rotated the subject in the corresponding direction. In the Mod condition, subjects were provided with a virtual miniature 3-D model of the space that could be manipulated using a mouse and monitor (similar in some ways to the WIM of Stoakley et al.). Subjects were able to rotate and view the model from any vantage point as well as zoom in closely to examine specific details. The graphical model of the space used for this condition was essentially identical to that used in the two virtual ''walk-through'' conditions VE and NVE that were mentioned earlier. The only differences were the point-of-view change and removal of the ceiling so that the internal space could be seen from the outside. The viewpoint for the Mod condition was controlled as follows: to translate the model up (down), click the right mouse button with the cursor at the top (bottom); to rotate the model left (right), click the middle mouse button with the cursor at the left (right); and to zoom in (out) on the model, click the middle mouse button with the cursor at the top (bottom).

3.4 Technical Details

The virtual environments in this study (VE, NV, and Mod) were run on modified Easyscene software from Coryphaeus running on an SGI Onyx RE / 2. The Onyx was equipped with 128 MB of RAM and two R4400 processors operating at 150 MHz. The Easyscene software was modified to allow for first-person navigation through the developed model, with collision detection and joystick support added. Modifications were also made to enable support for the HMD and adaptation of multiple viewpoints and control methods for the various experimental training conditions.

The HMD used in the VE condition was a Virtual Research VR4 (horizontal resolution 350 lines, vertical resolution 230 lines, 60 deg. field of view, weight 33 oz.). The HMD was not stereo-enabled for this study.[3] A Polhemus Fastrak provided orientation and position information on the HMD. The Polhemus positional information was discarded, and the image presented by the HMD was controlled by the Polhemus orientation vari-

3. Because of all the other depth cues available (such as those provided by perspective), it seemed unlikely that stereoscopy would have had much effect on performance in this experiment. See also the comments in Brooks (1992).

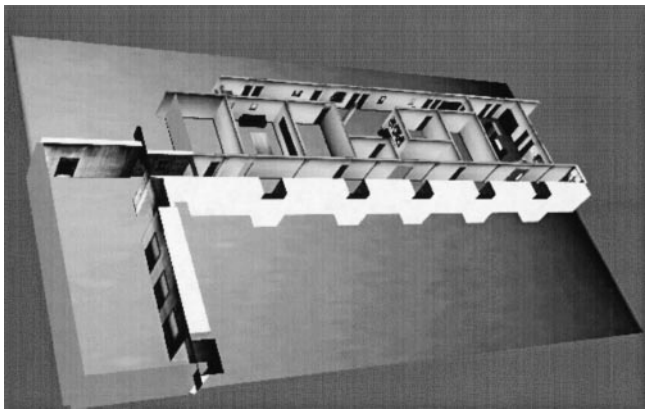

Figure 2. *Illustrative view of the space in Mod condition: far away.*

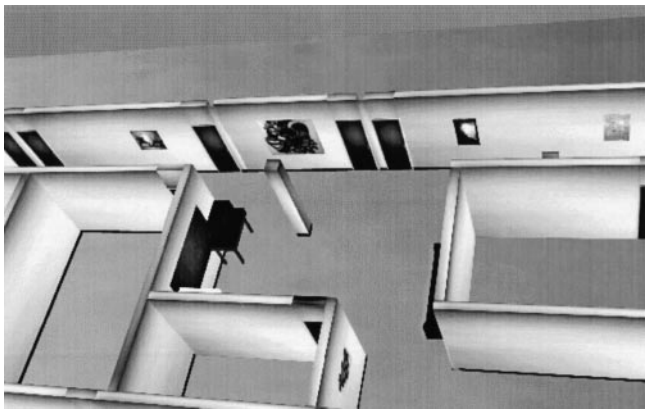

Figure 3. *Illustrative view of the space in Mod condition: zoomed in.*

ables and the joystick variables. In the NVE condition, the 21 in. monitor was operated in a 640 × 480 graphics mode and the image was controlled solely by the joystick. In the Mod condition, the same monitor was used; however, the joystick control was replaced by the mouse control.

An architectural model of the seventh floor of Building 36 was constructed from blueprints using Coryphaeus' Designers Workbench software. Accessible areas of the virtual environment were modeled with open doorways. Any closed doors in the virtual environment signified inaccessible areas irrelevant to the experiment. Motion of the subject was confined to the *x* and *y* plane. Each room in the model was populated with objects created from within Designer's Workbench, and all objects were fully texture mapped. Actual textures were obtained by using a Nikon N6006 camera with film scanned at 1280 × 1024 or by using a Kodak DC20 digital camera. Texture maps were edited using Kodak PhotoEasy software and Adobe Photoshop, and imported into Designer's Workbench for application to the seventh-floor model.

Figures 2 and 3 show illustrative views of the space obtained in the Mod training condition. Figure 4 shows a view of the space by the elevator doors obtained in the VE (or, equivalently, the NVE) condition. This same region of space can be seen in the lower-left corner of Figure 2.

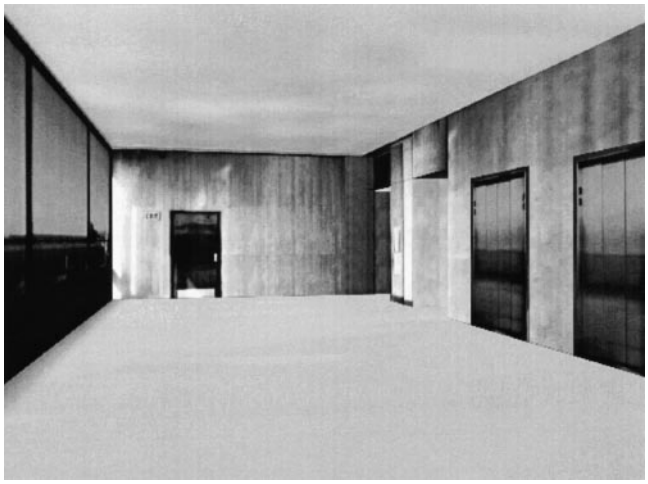

Figure 4. *Portion of space by elevators as seen in VE or NVE conditions (also pictured in lower-left corner from a different viewpoint in Figure 2).*

3.5 Subjects

Thirty-six MIT students were recruited for this study. Data from one subject were excluded due to apparent disorientation in all subtasks. Of the remaining 35, ten were trained using the RW condition, ten using the VE condition, seven using the NVE condition, and eight using the Mod condition. (The unequal numbers were not part of the experimental plan; they resulted from time constraints imposed on the duration of the experiment.)

4 Results

In this section, we present the main results of our experiments. Tests of statistical significance are presented in Appendix A.[4]

Figure 5 shows plots in the *x*,*y* plane of the estimated and actual locations of the landmarks relative to the stations for each of the thirteen subtasks specified in Figure 1 and Table 1. Each point represents the response of one subject; the different symbols differentiate the training groups. Figures 6 and 7 present histograms of the bearing and range errors segregated according to training method but pooled over subtasks and subjects to give roughly 100 responses per histogram.[5]

Because errors in range estimation tend to increase with range in a more or less consistent fashion (that is, the ratio of the range error to the actual range is relatively independent of the actual range, consistent with Weber's law), range data were examined in terms of the logarithm of the ratio of the estimated range to the actual range, so that a positive error corresponds to an overestimation of the range. For each of the two estimate components, bearing and range, we examined both the signed errors and the absolute values of these errors. Whereas in all cases the range errors are unitless (because they are expressed in terms of a ratio), the bearing errors are throughout the paper always expressed in degrees.

Figures 8 and 9 show contour plots of the signed errors as a function of both training condition / subject and subtask.

Table 2 (see pages 646-648) presents means and standard deviations. The entries in square brackets in each subsection of the table give the means and standard deviations corresponding to the pooled histograms shown in Figures 6 and 7. A detailed description of all the entries in this two-page table is provided in its legend.

The first result that is immediately evident from these data is that there are few, if any, major differences across the different training conditions (see the histograms in Figures 6 and 7 and the summary of the grand means and standard deviations corresponding to these histograms shown in Table 3.) There appears to be a slight overall bias in bearing estimate of roughly -2 deg. and in log range estimate of roughly -0.06 (corresponding to a range estimate that is only 85% of the true range). Results from the RW condition are worse than results from all the other conditions with respect to both mean errors and standard deviations, for both bearing errors and range errors, and for both signed and absolute errors. The differences that appear most pronounced are in the log range data: the mean error in the signed data is relatively large for both RW and Mod, the mean error in the absolute data is relatively large for RW, and the standard deviation of the RW data is relatively large for both the signed and absolute errors. It appears from the data shown in Figures 5 and 6 and Table 3 that the synthetic environments used for training are no worse than the real environment, and that there are no dramatic differences in the training effectiveness of the different synthetic environments. With respect to the latter of these statements, it should be noted that all the values of the sensitivity index d' corresponding to differences among training groups (difference in means divided by the average of the standard deviations) are less than 0.3 for all the bearing data, less than 0.4 for the signed log range data, and less that 0.6 for the absolute log range

4. We have relegated the results of these tests to an appendix because we do not regard them as crucial to the discussion. Although we certainly want to avoid emphasizing results that are not statistically significant, we do not believe that results that are statistically significant are necessarily important. For example, a difference in the mean bearing error for the conditions VE and NVE might be statistically significant but fail to be of interest because the statistically significant difference in means may still be small relative to the standard deviations of the two groups. The fact that for the given number of trials the difference between the means is large enough relative to these standard deviations to make the difference statistically significant does not imply that it is large enough to produce a value of the sensitivity index d' that reaches a usefully defined difference threshold. Stated differently, the fact that the difference in the means is statistically significant does not imply that knowing to which population an item belongs produces a significantly positive value of information transfer. Also, of course, a statistically significant difference in the means between the VE and NVE groups might fail to be of interest because it is small relative to other types of differences between VE and NVE (for example, related to comfort or cost).

5. Two points should be noted about these histograms. First, in order to take account of the fact that the total number of trials for each training method was not the same, we have plotted the percentage of responses in a bin rather than the number of responses in the bin. Second, slightly more data are included in these histograms than in the *x*,*y* plots (less than 2% more) because the *x*,*y* plots show data only when estimates of both range and bearing were available and only when the point fell within the *x*,*y* region displayed.

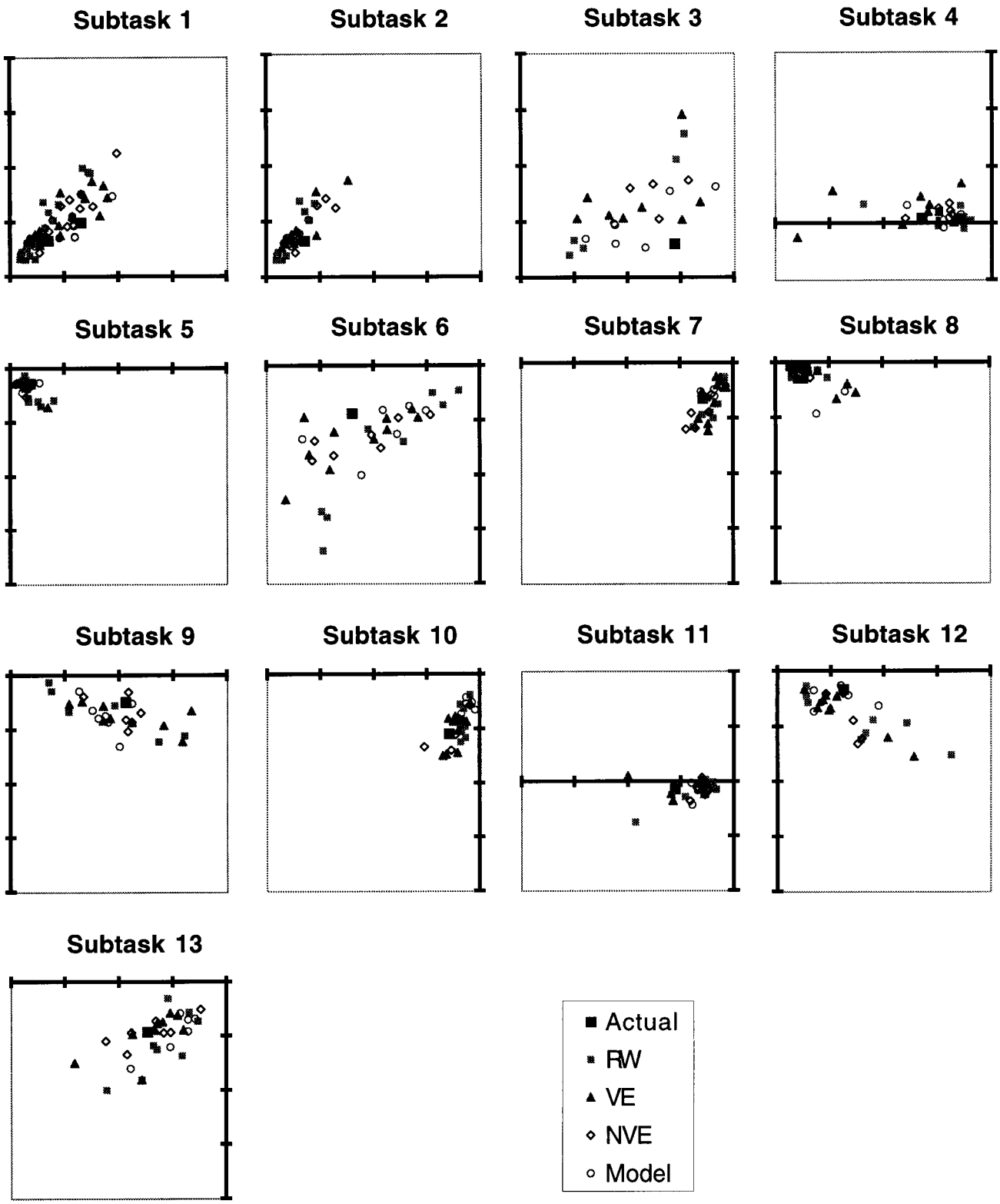

Figure 5. *Plots in the x,y plane of the estimated and actual locations of the targets relative to the reference locations for each of the subtasks 1 through 13 specified in Figure 1 and Table 1. The different training groups are indicated by different symbols. The distance between adjacent tick marks on both the x and y axes is 50 ft. The origin (0,0) is located at the intersection of the lines with tick marks. In order to maintain the same coordinate scale for all plots, a few points (less than 2%) exceeded the coordinate ranges used in these plots and are not shown. (This occurred only in the plots for subtasks 1, 9, and 12.)*

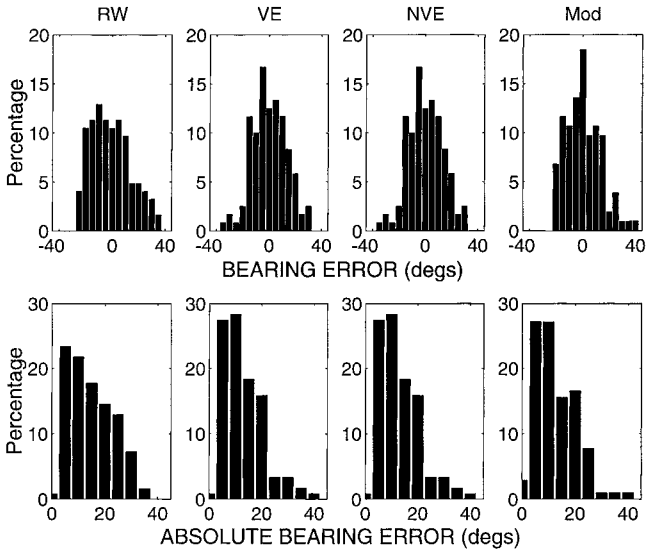

Figure 6. *Histograms of bearing errors. Top row shows signed errors, bottom row absolute errors. Different columns correspond to different training conditions.*

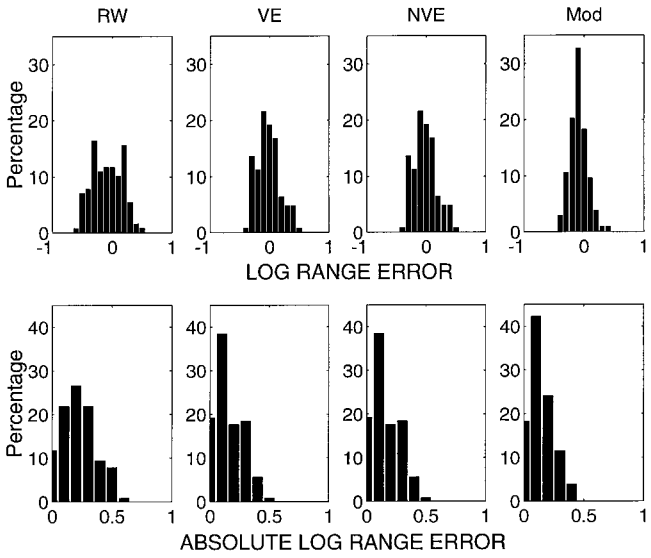

Figure 7. *Same as Figure 6, except for log range errors.*

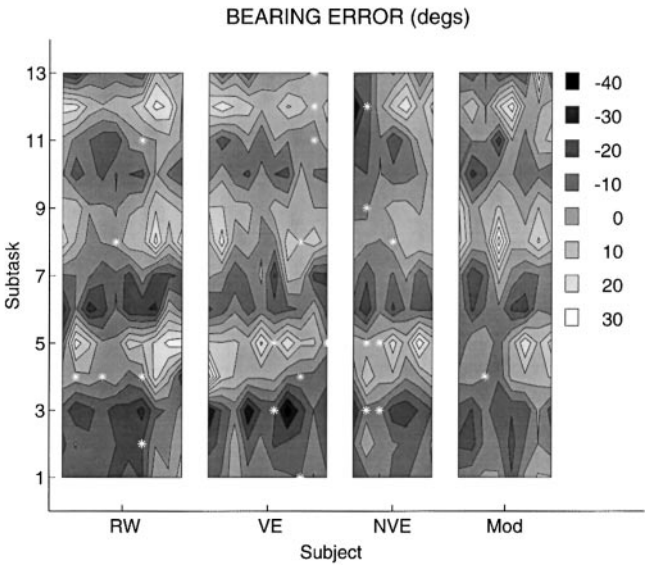

Figure 8. *Contour plot of signed bearing errors. Meaning of gray scale in terms of signed bearing errors is specified by column of numbers on right (in degrees). Contours are estimated by using two-dimensional linear interpolations between measured data points for grid of subject/condition (abcissa) and subtask (ordinate). Outliers and missing data points are plotted using white asterisks. Dominant effect of subtask is evident from primarily horizontal orientation of constant-error contours.*

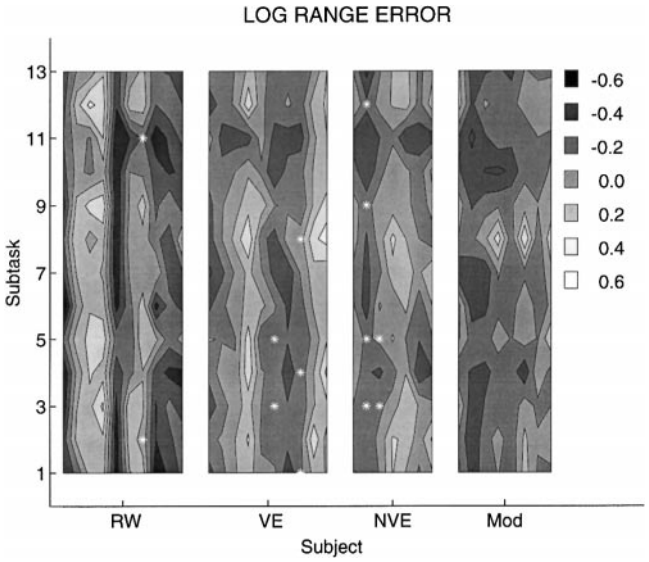

Figure 9. *Same as Figure 8 except for log range errors. Dominant effect of subject is evident from primarily vertical orientation of constant-error contours.*

data (the largest value being obtained for RW versus NVE in the absolute log range data).

A second important result to note is that the variance in the overall pooled results cannot be interpreted simply as repeated-measures variability. In other words, it is not

Table 2. *Means and Standard Deviations*

	Signed Errors						Absolute Errors				
	Subt / Subj (a)	Across Subjects Mean (b)	Across Subjects StDev (c)	Across Subtasks Mean (d)	Across Subtasks StDev (e)		Subt / Subj (a)	Across Subjects Mean (b)	Across Subjects StDev (c)	Across Subtasks Mean (d)	Across Subtasks StDev (e)
BEARING ERROR (DEGREES)											
RW	1	−10.20	10.08	−4.92	15.72		1	12.60	6.38	13.54	8.66
[−2.89	2	−11.33	9.01	−4.50	16.80	[12.76	2	12.22	7.60	14.83	8.02
14.98]	3	−16.05	8.47	−7.31	11.94	8.28]	3	16.25	8.04	11.62	7.37
	4	3.07	12.18	−9.58	8.20		4	9.36	7.56	10.58	6.72
	5	14.00	14.40	−7.83	10.35		5	16.60	10.91	10.50	7.33
	6	−17.40	11.45	−6.62	14.07		6	18.20	9.98	13.23	7.47
	7	−13.90	5.49	−5.80	16.63		7	13.90	5.49	15.20	7.57
	8	11.17	12.14	7.15	18.70		8	12.72	10.28	16.77	10.02
	9	6.70	7.54	4.88	13.61		9	7.90	6.12	10.73	9.29
	10	−6.70	6.27	3.92	14.36		10	7.90	4.46	11.00	9.58
	11	−2.78	9.48				11	7.67	5.70		
	12	16.30	9.26				12	16.30	9.26		
	13	−8.10	12.97				13	12.70	7.86		
	mean	−2.71	9.90	−3.06	14.04		mean	12.64	7.66	12.80	8.20
VE	1	−6.56	4.64	0.00	20.41		1	7.22	3.35	14.92	13.24
[−1.25	2	−12.50	7.37	0.15	16.31	[11.10	2	13.30	5.60	14.46	6.29
13.74]	3	−21.11	10.65	1.23	9.98	8.14]	3	21.11	10.65	8.00	5.64
	4	8.89	16.37	−4.62	13.65		4	12.44	13.52	11.85	7.58
	5	16.75	7.17	−0.54	11.10		5	16.75	7.17	7.77	7.63
	6	−10.30	6.58	−4.45	12.26		6	11.10	4.93	10.64	6.90
	7	−8.40	11.85	−0.92	17.23		7	11.40	8.63	13.08	10.60
	8	7.22	10.35	0.70	12.45		8	9.44	8.09	10.10	6.51
	9	7.20	4.42	−0.60	11.41		9	7.60	3.60	8.80	6.68
	10	−4.30	6.77	−3.50	11.41		10	6.90	3.67	10.50	4.81
	11	−2.33	7.75				11	6.33	4.58		
	12	16.56	5.50				12	16.56	5.50		
	13	−2.89	8.01				13	5.33	6.44		
	mean	−0.91	8.26	−1.25	13.62		mean	11.19	6.59	11.01	7.59
NVE	1	−8.71	5.56	−9.92	11.40		1	8.71	5.56	11.77	9.30
[−2.20	2	−6.43	5.88	−7.33	13.12	[10.43	2	7.57	3.99	12.67	7.23
12.61]	3	−19.80	6.42	−0.82	4.45	7.33]	3	19.80	6.42	3.91	1.92
	4	7.64	7.64	−3.42	14.94		4	8.21	6.92	13.42	6.29
	5	18.60	10.06	−0.85	12.84		5	18.60	10.06	10.08	7.45
	6	−13.00	6.58	−0.12	13.89		6	13.00	6.58	10.81	8.14
	7	−8.64	5.51	5.62	10.76		7	8.64	5.51	10.31	5.89
	8	5.58	7.05				8	6.92	5.44		
	9	4.92	6.04				9	6.58	3.67		
	10	−3.43	6.58				10	6.29	3.30		
	11	−5.00	11.82				11	11.00	5.26		
	12	7.42	21.54				12	18.42	11.09		
	13	−3.29	7.91				13	7.00	4.24		
	mean	−1.86	8.35	−2.41	11.63		mean	10.83	6.00	10.42	6.60

Table 2. *(Continued)*

	Signed Errors						Absolute Errors					
	Subt / Subj (a)	Across Subjects Mean (b)	Across Subjects StDev (c)	Across Subtasks Mean (d)	Across Subtasks StDev (e)		Subt / Subj (a)	Across Subjects Mean (b)	Across Subjects StDev (c)	Across Subtasks Mean (d)	Across Subtasks StDev (e)	
Mod	1	−2.44	7.33	−2.31	12.03		1	5.31	5.31	9.38	7.42	
[−1.42	2	−12.06	6.95	−7.85	12.23	[10.6	2	12.06	6.95	11.38	8.71	
13.06]	3	−12.69	5.93	−3.50	7.37	7.59]	3	12.69	5.93	6.50	4.64	
	4	4.93	7.72	−0.38	17.02		4	7.50	4.72	13.00	10.34	
	5	7.50	12.99	−1.62	15.33		5	11.75	8.68	12.54	8.21	
	6	−10.75	8.07	−3.08	13.59		6	10.75	8.07	10.92	8.11	
	7	−9.38	9.24	5.88	10.20		7	11.38	6.16	9.04	7.28	
	8	14.94	12.03	1.35	13.12		8	14.94	12.03	11.73	4.99	
	9	9.06	6.73				9	9.06	6.73			
	10	−6.75	7.19				10	7.50	6.28			
	11	−1.00	11.34				11	8.50	6.87			
	12	10.00	11.25				12	10.75	10.43			
	13	−9.00	14.95				13	15.25	7.09			
	mean	−1.36	9.36	−1.44	12.61		mean	10.57	7.33	10.56	7.46	
LOG RANGE ERROR												
RW	1	−0.13	0.26	−0.34	0.13		1	0.25	0.13	0.34	0.13	
[−0.09	2	−0.08	0.26	0.06	0.12	[0.22	2	0.23	0.11	0.11	0.08	
0.25]	3	−0.06	0.31	0.17	0.16	0.14]	3	0.27	0.13	0.19	0.12	
	4	−0.15	0.27	0.23	0.12		4	0.26	0.15	0.23	0.12	
	5	0.06	0.21	−0.39	0.11		5	0.18	0.09	0.39	0.11	
	6	−0.13	0.29	−0.01	0.16		6	0.26	0.18	0.12	0.10	
	7	−0.09	0.22	0.11	0.10		7	0.18	0.14	0.14	0.07	
	8	−0.06	0.23	−0.24	0.19		8	0.18	0.15	0.28	0.12	
	9	−0.05	0.30	−0.23	0.11		9	0.25	0.15	0.23	0.11	
	10	−0.19	0.17	−0.19	0.14		10	0.19	0.16	0.21	0.10	
	11	−0.19	0.24				11	0.26	0.16			
	12	0.02	0.27				12	0.24	0.10			
	13	−0.09	0.19				13	0.17	0.11			
	mean	−0.09	0.25	−0.08	0.13		mean	0.22	0.14	0.22	0.11	
VE	1	−0.02	0.17	−0.19	0.12		1	0.13	0.10	0.20	0.12	
[−0.02	2	0.01	0.18	−0.11	0.08	[0.16	2	0.14	0.11	0.11	0.08	
0.19]	3	−0.04	0.17	0.00	0.10	0.11]	3	0.15	0.08	0.08	0.07	
	4	0.04	0.23	0.19	0.18		4	0.17	0.15	0.23	0.11	
	5	−0.02	0.19	−0.02	0.11		5	0.15	0.11	0.08	0.06	
	6	0.00	0.16	−0.17	0.12		6	0.14	0.07	0.17	0.12	
	7	−0.08	0.21	−0.17	0.11		7	0.18	0.13	0.18	0.10	
	8	0.07	0.29	−0.14	0.07		8	0.24	0.15	0.14	0.07	
	9	0.03	0.19	0.15	0.12		9	0.15	0.10	0.15	0.11	
	10	−0.07	0.14	0.24	0.13		10	0.13	0.08	0.24	0.13	
	11	−0.12	0.20				11	0.20	0.11			
	12	0.01	0.21				12	0.16	0.13			
	13	−0.03	0.14				13	0.12	0.08			
	mean	−0.02	0.19	−0.02	0.11		mean	0.16	0.11	0.16	0.10	

Table 2. *(Continued)*

	Signed Errors						Absolute Errors					
	Subt / Subj (a)	Across Subjects Mean (b)	Across Subjects StDev (c)	Across Subtasks Mean (d)	Across Subtasks StDev (e)		Subt / Subj (a)	Across Subjects Mean (b)	Across Subjects StDev (c)	Across Subtasks Mean (d)	Across Subtasks StDev (e)	
NVE	1	−0.01	0.16	−0.05	0.11		1	0.13	0.08	0.09	0.08	
[−0.03	2	0.08	0.17	−0.26	0.09	[0.14	2	0.14	0.11	0.26	0.09	
0.17]	3	0.05	0.11	−0.05	0.11	0.10]	3	0.08	0.09	0.09	0.07	
	4	−0.13	0.16	0.10	0.13		4	0.18	0.09	0.13	0.10	
	5	−0.03	0.12	0.08	0.15		5	0.09	0.06	0.14	0.09	
	6	−0.02	0.16	−0.11	0.13		6	0.12	0.09	0.14	0.09	
	7	0.03	0.18	0.03	0.17		7	0.15	0.07	0.14	0.09	
	8	0.00	0.17				8	0.12	0.10			
	9	−0.01	0.09				9	0.06	0.07			
	10	−0.07	0.16				10	0.14	0.09			
	11	−0.26	0.09				11	0.26	0.09			
	12	0.08	0.12				12	0.12	0.07			
	13	−0.04	0.19				13	0.14	0.12			
	mean	−0.02	0.14	−0.04	0.13		mean	0.13	0.09	0.14	0.09	
Mod	1	−0.08	0.13	−0.02	0.16		1	0.12	0.08	0.14	0.08	
[−0.09	2	−0.07	0.14	−0.25	0.08	[0.15	2	0.11	0.10	0.25	0.08	
0.15]	3	−0.10	0.14	−0.17	0.11	0.10]	3	0.15	0.09	0.18	0.10	
	4	−0.15	0.12	−0.07	0.16		4	0.18	0.08	0.13	0.11	
	5	0.01	0.13	−0.15	0.07		5	0.10	0.07	0.15	0.07	
	6	−0.08	0.14	0.01	0.18		6	0.14	0.08	0.13	0.12	
	7	−0.16	0.15	−0.07	0.08		7	0.18	0.12	0.09	0.05	
	8	0.07	0.22	0.01	0.11		8	0.18	0.13	0.08	0.07	
	9	−0.06	0.08				9	0.07	0.07			
	10	−0.23	0.13				10	0.23	0.13			
	11	−0.17	0.11				11	0.17	0.11			
	12	−0.01	0.13				12	0.11	0.07			
	13	−0.13	0.14				13	0.16	0.09			
	mean	−0.09	0.14	−0.09	0.12		mean	0.15	0.09	0.15	0.09	

The first section of this table gives the results for bearing errors, the second for log range errors (log to the base ten of the ratio of estimated range to actual range). The four blocks of rows in each section give the results for the four training conditions RW, VE, NVE, and Mod. The two blocks of columns in each section give the results for the signed errors and for the absolute errors.

In each of the sixteen subtables, the two numbers in square brackets give the grand mean and the grand standard deviation corresponding to that subtable, that is, the statistics that are obtained when the data are pooled over all subjects and subtasks for the given training condition and the given type of error (signed or absolute) associated with that subtable. (In all cases, the term *standard deviation* refers to the standard deviation about the mean.)

Within each of the sixteen subtables, column (a) serves to identify both the subtask (see Table 1) and the subject; columns (b) and (c) give the means and standard deviations, respectively, for each subtask (where the variable averaged across is the subject); and columns (d) and (e) give the means and standard deviations for each subject (where the variable averaged across is subtask). The row labeled *mean* in each subtable gives the mean values for columns (b), (c), (d), and (e).

the case that the scatter in the data can be explained by assuming a single underlying random variable with a probability density that is blind to differences across subtasks and subjects.

The most obvious structural element in the data that contradicts such a simple model is the remarkably strong dependence of signed bearing error on subtask. This dependence is evident in the *x*,*y* plots shown in Figure 5,

Table 3. *Grand Means and Standard Deviations*

	Bearing		Log Range	
	Signed	Absolute	Signed	Absolute
RW	−2.9, 15.0	12.8, 8.3	−0.09, 0.25	0.22, 0.14
VE	−1.3, 13.7	11.1, 8.1	−0.02, 0.19	0.16, 0.11
NVE	−2.2, 12.6	10.4, 7.3	−0.03, 0.17	0.14, 0.10
Mod	−1.4, 13.1	10.6, 7.6	−0.09, 0.15	0.15, 0.10
Mean	−2.0, 13.6	11.2, 7.8	−0.06, 0.19	0.17, 0.11

Table 4. *Correlation ρ Between Pairs of Training Groups for the Dependence of Mean Errors (Across Subjects) on Subtasks*

	Bearing		Log Range	
	Signed	Absolute	Signed	Absolute
ρ (RW, VE)	0.95	0.65	0.39	−0.09
ρ (RW, NVE)	0.91	0.70	0.61	0.10
ρ (RW, Mod)	0.94	0.45	0.76	−0.30
ρ (VE, NVE)	0.96	0.82	0.42	0.32
ρ (VE, Mod)	0.89	0.25	0.69	0.26
ρ (NVE, Mod)	0.82	0.16	0.47	0.51
Mean	0.91	0.50	0.56	0.13

in the contour plots shown in Figures 8 and 9, and in the results listed in Table 2. For example, whereas there is a large negative bearing error on subtask 3, there is a large positive error on subtask 5. According to the results listed in Table 2, the mean signed bearing errors on subtask 3 for the four training groups are −16.1, −21.1, −19.8, and −12.7, whereas the comparable results for subtask 5 are 14.0, 16.8, 18.6, and 7.5. On the average (across training groups), the means and standard deviations for these two tasks are m(task 3) = −17.3 deg., m(task 5) = 14.2 deg., σ (task 3) = 7.9 deg., and σ (task 5) = 11.2 deg. Using these averages to make a crude estimate of d', one obtains a value of d' (task 3, task 5) greater than 3. Obviously, there is a strong effect of subtask in the signed bearing errors.

Further results concerning the role played by subtask are presented in Tables 4 and 5. Table 4 shows the correlation coefficients between pairs of training groups for the dependence of mean errors (averaged across subjects in the group) on subtask, for each of the four error types (signed bearing error, absolute bearing error, signed log range error, and absolute log range error). The values of the correlation coefficient ρ for the singed bearing errors ($0.82 \leq \rho \leq 0.96$; Ave ρ = 0.91) are remarkably high. Apparently, something about the structure of the subtasks played a very powerful role in determining the values of the signed bearing errors. The absolute bearing errors and the signed log range errors also show some substantial correlations, but they are not nearly as pronounced.

Table 5 compares the overall standard deviations (shown inside the square brackets in Table 2) to the standard deviations arising from variations across subjects (the mean value of column (c) in Table 2) and the standard deviations arising from variations across subtasks (the mean value of column (e) in Table 2). As can be seen in the results for signed bearing errors, most of the variation in the pooled results (characterized by the grand σ) is the result of variation across subtasks. For example, in the RW condition, for which the grand σ is 15.0, the σ associated with variation across subtasks is 14.0, whereas the σ associated with variation across subjects is only 9.9. This is also true for the absolute bearing error, but to a much smaller extent. As was evident in the above correlation analysis (Table 4), it is also clear from this table that subtask plays a smaller role in range estimates than in bearing estimates. A way of describingthe strong dependence of signed bearing error on subtask is outlined briefly in Section 5.

It is also clear that the results depend strongly on the subject. Because the tests with the different training methods used different subjects, one cannot examine the correlation between methods for the dependence of errors (averaged over subtasks) on subject. However, as would be expected on the basis of previous research in this area—and as can be seen by examining the means and standard deviations in Table 2—intersubject differences constitute a substantial source of overall variance.

The fact that variation among subjects plays an important role, particularly in the range estimates, is also evident in Table 5. In contrast to the manner in which the

Table 5. *Comparisons of Grand Standard Deviations to Standard Deviations Across Subjects and Across Subtasks*

	Bearing		Log Range	
	Signed	Absolute	Signed	Absolute
Grand σ				
RW	15.0	8.3	0.25	0.14
VE	13.7	8.1	0.19	0.11
NVE	12.6	7.3	0.17	0.10
Mod	13.1	7.6	0.15	0.10
Mean	13.6	7.8	0.19	0.11
X subj σ				
RW	9.9	7.6	0.25	0.14
VE	8.3	6.6	0.19	0.11
NVE	8.4	6.0	0.14	0.09
Mod	9.4	7.3	0.14	0.09
Mean	9.0	6.9	0.18	0.11
X subt σ				
RW	14.0	8.2	0.13	0.12
VE	13.6	7.6	0.11	0.10
NVE	11.6	6.6	0.13	0.09
Mod	12.6	7.5	0.12	0.09
Mean	12.9	7.5	0.12	0.10

Table 6. *Values of σ(m)/m(σ)*

	Bearing		Log Range	
	Signed	Absolute	Signed	Absolute
Subtasks				
RW	1.18	0.47	0.30	0.28
VE	1.39	0.72	0.28	0.30
NVE	1.23	0.83	0.63	0.55
Mod	1.03	0.40	0.61	0.46
Mean	1.21	0.61	0.30	0.40
Subjects				
RW	0.43	0.27	1.66	0.86
VE	0.16	0.33	1.42	0.58
NVE	0.44	0.47	0.96	0.65
Mod	0.32	0.29	0.79	0.62
Mean	0.34	0.34	1.21	0.68

intersubtask standard deviation (*x* subt σ) is the main contributor to the grand standard deviation (Grand σ) for the signed bearing errors, the intersubject standard deviation (*x* subj σ) is the main contributor to the grand variance for the signed log range errors.

The manner in which the bearing and range data relate to the variations across subtasks and subjects is further displayed in Table 6.

In the upper portion of this table, we have tabulated the values of $\sigma(m)/m(\sigma)$, where $\sigma(m)$ denotes the standard deviation of the means shown in column (b) of Table 2, and $m(\sigma)$ denotes the mean of the standard deviations shown in column (c) of Table 2. Large values of $\sigma(m)/m(\sigma)$ correspond to the case in which the differences among different subtasks are large relative to variation across subjects. In the lower portion of the table, wehave tabulated the values of $\sigma(m)/m(\sigma)$, where $\sigma(m)$ denotes the standard deviation of the means shown in column (d) of Table 2, and $m(\sigma)$ denotes the mean of the standard deviations shown in column (e) of Table 2. Here, large values of $\sigma(m)/m(\sigma)$ correspond to the case in which the differences among different subjects are large relative to the variations across subtasks. Loosely speaking, the values of $\sigma(m)/m(\sigma)$ in the top portion of the table can be interpreted as average values of d' between pairs of subtasks, and, in the bottom portion, as average values of d' between pairs of subjects. As can be seen from Table 6 as well as from Table 5, intersubtask differences are more pronounced for bearing errors, whereas intersubject differences are more pronounced for range errors. For example, confining attention to signed errors, one finds that the mean d' for bearing errors is roughly four times greater between subtasks than between subjects (1.21 compared to 0.34), whereas for log range errors it is roughly four times smaller between subtasks than between subjects (0.30 compared to 1.21). The same phenomenon is evident graphically in Figures 8 and 9: the striations in the bearing errors are primarily horizontal, whereas the striations in the range errors are primarily vertical.

In addition to the above observations, it is important to note that, for a given subject, the correlation between bearing errors and range errors is relatively weak. For the

case of signed errors, the correlation ρ lies in the range $-0.56 \leq \rho \leq 0.80$ for the 35 subjects (Ave = 0.22, $\sigma = 0.32$). For the case of absolute errors, the range is $-0.60 \leq \rho \leq 0.65$ (Ave = -0.01, $\sigma = 0.35$). Although for any one subject the correlation may be significant, across subjects there is no systematic trend in the relationship between either signed bearing errors and signed log range errors or between absolute bearing errors and absolute log range errors.

As indicated in Section 3, we did not in this preliminary experiment pay attention to the manner in which subjects were assigned to groups, nor did we administer any spatial-abilities tests to the selected subjects. Thus, we are not in a position either to interpret intersubject differences or to guarantee that our results are not biased by the composition of the various training groups.

A cursory examination of the effects of gender in the RW and Mod groups (the first of which contained six males and four females, the second of which contained four males and four females) showed no pronounced gender effect in either the means or the variances, except possibly for the mean results in the signed log range data which suggested a possible tendency for females to underestimate distance to a slightly greater degree than males. This result is consistent with the difference between these two groups and the VE and NVE groups (each of which contained only one female) with respect to the mean signed log range errors shown in Table 3. Whereas the first two groups had a mean error of -0.09, the last two had mean errors of -0.02 and -0.03. (In both the RW and Mod groups, the mean log range error for the males was -0.06 and the mean log range error for the females was -0.12.) Obviously, serious study of individual differences related to gender or any other subject descriptor will require extensive pretesting of spatial abilities and systematic subject selection, as well as an increased number of subjects.[6]

6. We have included these comments on possible gender effects only because we know that, if we had omitted them, many readers would have wondered about the gender composition of the groups and how it related to the results. The study was not designed, and the results cannot be used, to evaluate gender effects in a statistically meaningful manner.

Finally, in considering the above results, it should be noted that, before these data were collected, each subject was required to perform the task of estimating the bearing and range of a landmark that was visible from the station (the water fountain from station A). As one might expect, the bias in signed bearing estimate for this visible-landmark task tended to be very small: the mean estimated bearing for each of the training groups differed from the actual bearing (23 deg.) by less than 1 deg. In contrast, however, the bias in the range estimate was surprisingly large: the mean signed log range error for the groups was -0.20 (RW), -0.11 (VE), -0.08 (NVE), and -0.21 (Mod). Not only do these errors seem large on an absolute basis (for example, -0.21 corresponds to a range estimate that is only about 60% of the actual range of 25 feet), but they are large relative to the mean errors encountered in the tests analyzed above in which the landmarks were always hidden from view and the estimates depended on the subject's memory. (The corresponding figures for these tests according to the results shown in Table 3 were -0.09 (RW), -0.02 (VE), -0.03 (NVE), and -0.09 (Mod).) This suggests, perhaps, that the subjects had an inflated picture of the length corresponding to one foot, and, with respect to this inflated unit of length, that they tended to overestimate the range of the landmarks that were not visible. Alternatively, one might simply conclude that landmarks tend to be judged as nearer when they are visible than when they are hidden from view.

In order to determine the extent to which the estimates in this visible-target task (which we refer to as subtask 0) were correlated on a subject-by-subject basis with the mean value of the estimates for the invisible-target tasks 1–13, correlation coefficients were computed for both bearing and log range, for both signed and absolute errors, and for both the overall pool of subjects and the individual training groups RW, VE, NVE, and Mod. The results of these computations (see Table 7) show that the correlations are relatively high for the RW log range data (both signed and absolute) and, to a slightly lesser extent, for the Mod signed bearing data and the log range data (both signed and absolute). The case that exhibits the lowest correlation is the absolute bearing errors.

Table 7. *Correlations with Subtask 0 (Target Visible)*

	Bearing		Log Range	
	Signed	Absolute	Signed	Absolute
RW	0.27	0.17	0.75	0.76
VE	0.40	0.16	0.19	0.35
NVE	0.49	0.01	0.28	0.46
Mod	0.68	−0.09	0.66	0.52
Mean	0.46	0.06	0.47	0.52
Pooled	0.42	−0.09	0.58	0.62

Table 8. *Results of Normalizing Data According to Estimates Made in Subtask 0*

	Bearing				Log Range			
	Signed		Absolute		Signed		Absolute	
	m	σ	m	σ	m	σ	m	σ
RW	−2.9	15.0	12.8	8.3	−0.09	0.25	0.22	0.14
RW*	−2.5	14.9	12.5	8.5	0.11	0.20	0.18	0.13
VE	−1.2	13.7	11.1	8.2	−0.02	0.19	0.16	0.11
VE*	−1.7	13.9	11.1	8.5	0.10	0.21	0.16	0.16
NVE	−2.2	12.7	10.4	7.3	−0.03	0.17	0.14	0.10
NVE*	−2.5	12.4	10.2	7.5	0.05	0.18	0.15	0.11
Mod	−1.4	13.1	10.6	7.7	−0.09	0.15	0.15	0.10
Mod*	−1.1	13.0	10.1	8.2	0.13	0.14	0.15	0.11
Mean	−1.9	13.6	11.2	7.9	−0.06	0.19	0.17	0.11
Mean*	−1.9	13.5	11.0	8.2	0.05	0.18	0.16	0.13

*Denotes results for normalized data.

In order to determine the extent to which the results would have changed had we normalized them by subtracting out the errors made on subtask 0 (on a subject-by-subject basis), we recomputed the means and standard deviations for each of the cases RW, VE, NVE, and Mod, for both the signed and absolute versions of these errors. The results of these further computations are shown in Table 8.

Overall, the changes due to normalization do not seem large. The change from negative to positive results for the means of the signed log range data is consistent with the fact, mentioned previously, that the underestimation of distance was exceptionally large for subtask 0. Also, the reduction in the standard deviation for RW in the signed log range case appears consistent with the large positive correlation for this case shown in Table 7. It is interesting to note, however, that large correlations in Table 6 for the RW absolute log range case and for both the signed and absolute versions of the Mod log range case are not accompanied by significant changes in σ when the data are normalized. It should also be noted that normalization increases σ in a number of cases (for example, the VE absolute log range case). In general, none of the main results previously discussed appear to be substantially altered by the normalization process.

5 Discussion

In general, it is clear from the results summarized in Section 4 that differences among the training methods considered are trivial in comparison to the variability of the data resulting from intersubtask and intersubject differences. Training with the use of a virtual environment appears to be as effective as training with the real environment, and there appear to be no dramatic differences among the results obtained with the different virtual environment systems.

Although we do not find these general results surprising in the light of the previous work cited in Section 2, it is difficult to make meaningful detailed comparisons with these previous studies because of the many important differences between these studies and our own. Not only were relatively few of the previous studies concerned with training transfer from the virtual environment to the real world, but, even within this small group of studies, the training test conditions were radically different from our own. For example, in the studies by Witmer et al. (1995), Witmer et al. (1996), Bliss et al. (1997), and Tate et al. (1997), the focus was on the acquisition of route knowledge rather than configurational knowledge.

If we had been forced to rate the training conditions in advance based on our intuition, we would have guessed that RW would have been best, NVE worst, and VE and Mod intermediate. We would have guessed that VE would be superior to NVE because of the often-cited

advantages of immersion; we would have guessed that Mod would do relatively well because of its similarity to a 3-D map (and the work by Stoakley et al. (1995)); and we would have guessed that RW would have been best, given the technical limitations of current simulation systems. Eventually, we expect that a combination of a good VE system and a good Mod system (and an appropriate method for transitioning between the two) will prove substantially superior to RW, and that this superiority will be evident for a much wider variety of test tasks than those considered in our preliminary experiment.

Additional results concern the nature of the response errors. Obviously, it would be worthwhile developing methods to eliminate them. Furthermore, to the extent that they cannot be eliminated, it would be worthwhile building models to predict them. In general, such efforts will serve both to improve spatial behavior and to improve our ability to distinguish between different training methods by reducing variability.

As discussed in Section 4, bearing errors were dominated by the subtask, more or less independent of subject and training method. Furthermore (see Figure 5), the bias in bearing estimation can be summarized by stating that subjects tended to ''square the rectangle'' (subjects thought that the rectangular space was wider than it actually was). Similarly consistent biases in bearing estimates have been noted previously (for example, see Henry (1992)).

Also, as shown in Section 4, range errors usually consisted of underestimates, and, unlike the bearing errors, the variability in these estimates was dominated by intersubject differences, not by intersubtask differences. Although underestimation of range was to be expected on the basis of previous work (for example, see Henry (1992) and Witmer and Kline (1998)), we were surprised that the RW group showed the most pronounced bias.

These biases in the bearing and range estimates are clearly of interest from the general viewpoint of the psychology of spatial perception. However, we suspect that they will prove to be relatively unimportant from the viewpoint of VE-assisted spatial training, because we believe that it will be possible to train subjects to eliminate these biases. In particular, we see no reason why a subject who consistently underestimates range cannot learn to achieve better accuracy by correct-answer feedback training. Moreover, we believe that effecting this correction in one spatial environment would transfer to most other environments. To the extent that this notion is correct, and to the extent that attention is focused on response bias, it matters relatively little whether the distance response chosen for the experiment is phrased in terms of ''estimated feet to the target'' or in terms of something that is more ecologically valid. If the bias is easy to eliminate, the source of the bias is relatively unimportant.

Bias in the bearing estimates may be less easy to eliminate, because it appears to be less homogeneous and more dependent on the subtask. In order to correct bias, it must be possible for the subject to develop a relatively simple mental model of how the bias depends on the situation. Whereas in the range case it appears that much of the bias could be removed if the subject merely learned to increase the range estimate by a fixed percentage, no such simple correction rule has yet been demonstrated for the bearing case. Whether or not it is possible to find a rule that is sufficiently simple to be of use in training subjects to eliminate the bearing bias in a wide range of spatial situations is a topic for future study.

Assuming that a subject's response bias in range and bearing can be eliminated or reduced by appropriate training, one is still faced with problems related to a subject's intrinsic response variability in estimating range and bearing, and to intersubject differences of various types. It may be that training can reduce intrasubject and / or intersubject response variability. However, even if it can't, it may be possible, through the use of appropriate spatial-abilities tests, to predict some aspects of a subject's performance in the experimental spatial-behavior tasks. To the extent that this is possible, it, too, like the elimination of bias, would enable one to sharpen one's ability to distinguish between different training methods and select the one most likely to be appropriate to the requirements of the specific training situation.

In conclusion, two points need to be stressed. First, to the extent that the issues related to response errors that have been discussed are independent of the choice among RW, VE, NVE, and Mod as training conditions, it doesn't matter which of these conditions are used in future work addressed to these issues. Second, our preliminary experiment was very limited in that only one,

relatively simple space was used; no intrasubject comparisons of training conditions were made; the training experience of the subjects was not varied systematically and in a controlled manner (other than the assignment to one of the conditions RW, VE, NVE, and Mod); and the range of tests used was highly constrained. With respect to this last point, it should be noted not only that attention was confined to configurational knowledge, but also that the tests of configurational knowledge were very limited. No map-drawing tests were included, and there were no attempts to explore a subject's ability to make spatial estimates that were not anchored at the location at which the subject was standing (for example, estimating the bearing and range to a target from a different location, or estimating whether or not a clear line of sight exists between two locations, neither of which coincided with the position of the subject). As recently stressed by Jim Templeman, Jack Loomis, and Rudy Darken (personal conversations), the inclusion of such responses would be of interest with respect to both basic research issues and VE training applications.

In future work concerned with training spatial behavior in specific spaces, we intend not only to study more complex spaces but also to develop methods for decreasing response bias and scatter in the experimental data, to include a wider variety of spatial behavior measures, to explore the possibility of creating a general VE spatial abilities test to help characterize the individual subjects used in the experimental studies, and to create and evaluate a VE training system that combines an immersive walk-through VE (with an improved motion interface) and a miniature-model VE in such a manner that the training provided by this system clearly exceeds that provided by real-world training.

Acknowledgments

This work was supported by the Office of Naval Research, Grant # N00014-96-1-0379. We are indebted to Terry Allard, Rudy Darken, Jack Loomis, and Jim Templeman for many useful discussions concerning the use of VE for training spatial behavior. We also wish to acknowledge many useful comments by the anonymous reviewers.

References

Aginsky, V., Harris, C., Rensink, R., & Beusman, J. (1996). *Two strategies for learning a route in a driving simulator.* (Technical report TR 96-6). Cambridge Basic Research Center, Nissan Research and Development.

Bliss, J. P., Tidwell, P. D., & Guest, M. A. (1997). The effectiveness of virtual reality for administering spatial navigation training to firefighters. *Presence: Teleoperators and Virtual Environments, 6*(1), 73–86.

Brooks, F. P., Jr. (1992). *Six generations of building walkthrough: Final technical report to the National Science Foundation* (Technical report TR92-026).

Chance, S. S., Gaunet, F., Beall, A. C., & Loomis, J. M. (1998). Locomotion mode affects the updating of objects encountered during travel: The contribution of vestibular and proprioceptive inputs to path integration. *Presence: Teleoperators and Virtual Environments, 7*(2), 168–178.

Chase, W. G. (1983). Spatial representation in taxi drivers. In D. R. Rogers & J. A. Sloboda (Eds.), *Acquisition of Symbolic Skills.* New York: Plenum.

Colle, H. A., & Reid, G. B. (1998). The room effect: Metric spatial knowledge of local and separated regions. *Presence: Teleoperators and Virtual Environments, 7*(2), 116–128.

Darken, R. P., & Banker, W. P. (1998). Navigating in natural environments: A virtual environment training transfer study. *Proceedings of VRAIS '98,* 12–19.

Darken, R. P., & Sibert, J. L. (1993). A toolset for navigation in virtual environments. *Proceedings of ACM User Interface Software & Technology,* 157–165.

———. (1996a). Navigating large virtual spaces. *International Journal of Human-Computer Interaction,* 49–71.

———. (1996b). Wayfinding strategies and behaviors in large virtual worlds. *Proceedings of Computer Human Interfaces Conference,* 1976 (CHI'96), 142–149.

Eckstrom R. B., French, J. W., Harmen, H. H., & Dermen, D. (1976). *Manual for the kit of factor-referenced cognitive tests.* Technical Report, Office of Naval Research Contract N00014-71-C-0117. Princeton, N.J.: Educational Testing Service.

Goerger, S. R. (1998). *Spatial knowledge acquisition and transfer from virtual to natural environments for dismounted land navigation.* Unpublished master's thesis, Department of Computer Science, Naval Postgraduate School.

Goerger, S. R., Darken, R. P., Boyd, M. A., Gagnon, T. A., Liles, S. W., Sullivan, J. A., & Lawson, J. P. (1998). Spatial knowledge acquisition from maps and virtual environments

in complex architectural spaces. *Proceedings of the 16th Applied Behavioral Sciences Symposium,* U.S. Air Force Academy, Colorado Springs, CO., 6–10.

Guilford, J. P., & Zimmerman, W. S. (1947). *The Guilford-Zimmerman Aptitude Survey.* Palo Alto: Consulting Psychologists Press.

Henry, D. P. (1992). *Spatial perception in virtual environments: Evaluating an architectural application*. MSEE thesis, University of Washington.

Loomis, J. M., Golledge, R. G., & Klatzky, R. L. (1998). Navigation system for the blind: Auditory display modes and guidance. *Presence: Teleoperators and Virtual Environments, 7*(2), 193–203.

May, M., Peruch, P., & Savoyant, A. (1995). Navigating in a virtual environment with map-acquired knowledge: Encoding and alignment effects. *Ecological Psychology, 7*(1), 21–36.

Moeser, S. D. (1988). Cognition mapping in a complex building. *Environment and Behavior, 20,* 21–49.

Pausch, R., Proffitt, D., & Williams, G. (1997). Quantifying immersion in virtual reality. *Computer Graphics Proceedings.* Annual Conference Series.

Peruch, P., Vercher, J.-L., & Gauthier, G. M. (1995). Acquisition of spatial knowledge through visual exploration of simulated environments. *Ecological Psychology, 7*(1), 1–20.

Presson, C. C., & Montello, D. R. (1994). Updating after rotational and translational body movements: Coordinate structure of perspective space. *Perception, 23,* 1447–1455.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1157–1165.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1997). Navigating buildings in 'Desk-top' Virtual Environments: Experimental investigations using extended navigational experience. *Journal of Experimental Psychology: Applied, 3*(2), 143–159.

———. (1998). Navigating large-scale 'Desk-Top' virtual buildings: effects of orientation aids and familiarity. *Presence: Teleoperators and Virtual Environments, 7*(2), 179–192.

———. (1999). Navigating large-scale virtual environments: What differences occur between helmet-mounted and desk-top displays. *Presence: Teleoperators and Virtual Environments, 8*(2), 157–168.

Ruddle, R. A., Randall, S. J., Payne, S. J., & Jones, D. M. (1996). Navigation and spatial knowledge acquisition in large-scale virtual buildings: An experimental comparison of immersive and desk-top displays. *Proceedings of the Second International FIVE Conference,* 125–136.

Satalich, G. A. (1995). *Navigation and wayfinding in virtual reality: Finding the proper tools and cues to enhance navigational awareness.* MSEE thesis, University of Washington.

Stoakley, R., Conway, M. J., & Pausch, R. (1995). Virtual reality on a WIM: Interactive worlds in miniature. *CHI'95 Mosaic of Creativity.*

Tate, D. L., Sibert, L., & King, T. (1997). Virtual environments for shipboard firefighting training. *Proceedings of IEEE Virtual Reality Annual International Symposium (VRAIS '97),* 61–68.

Thorndyke, P. W., & Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology, 14,* 560–589.

Tlauka, M., & Wilson, P. N. (1996). Orientation-free representations from navigation through a computer-simulated environment. *Environment and Behavior, 28*(5), 647–664.

Waller, D., Hunt, E., & Knapp, D. (1998). The transfer of spatial knowledge in virtual environment training. *Presence: Teleoperators and Virtual Environments, 7*(2), 129–143.

Witmer, B. G., Bailey, J. H., & Knerr, B. W. (1995). *Training dismounted soldiers in virtual environments: Route learning and transfer.* USARI Technical Report 1022.

Witmer, B. G., Bailey, J. H., Knerr, B. W., & Parsons, K. C. (1996). Virtual spaces and real world places: transfer of route knowledge. *International Journal of Human-Computer Studies, 45,* 413–428.

Witmer, B. G., & Kline, P. B. (1998). Judging perceived and traversed distance in virtual environments. *Presence: Teleoperators and Virtual Environments, 7*(2), 144–167.

Appendix A: Analysis of Variance of Navigation Errors

ANOVA analysis was performed in order to test the statistical significance of the trends that were observed in the data. Bearing errors, absolute bearing errors, log range errors, and absolute log range errors were each analyzed in an ANOVA analysis in which training and subtask were the main factors and the subject factor was nested within training. In the bearing and absolute bearing data, 432 data points were included in the analysis. In the range data analysis, 442 data points were included.

Bearing error depended significantly on subtask ($F(12, 385) = 45.93$, $p < 0.001$) and subject ($F(31, 385) = 3.14$, $p < 0.001$), but not on training method

($F(3, 385) = 1.10$, $p > 0.001$). This analysis confirms the observation that there was a significant dependence of the bearing error on the subtask and on individual subject, but not on the method of training. Post-hoc tests confirmed that, for some subtasks, bearing errors were more negative (to the left) than in other subtasks. The expected bearing errors ranged from −17.4 deg. (in subtask 3) to 14.6 deg. (in subtask 5). Scheffe post-hoc tests confirmed that many pairs of the subtasks yielded significantly different mean bearing errors. In general, most of the significant differences arose because responses on subtasks 2, 3, 6, and 7 tended to be to the left of the target compared to responses on subtasks 5, 8, and 12, which tended to be to the right of the target.

Results of analysis of absolute bearing error were similar; however, subject differences were not as large. Absolute bearing error depended significantly on subtask ($F(12, 385) = 6.29$, $p < 0.001$), but not on training method ($F(3, 385) = 2.12$, $p > 0.001$) or on subject ($F(31, 385) = 1.38$, $p > 0.001$). Scheffe post-hoc analysis indicated that the expected absolute bearing errors were marginally significantly larger in subtask 3 (expected absolute error of 17.5 deg.) than in subtask 10 (expected absolute bearing error of 7.1 deg.).

The subtasks for which expected bearing error had the largest magnitude were the same subtasks for which the expected absolute errors were largest. Specifically, subtasks 2, 3, 5, 6, 7, 8, and 12 had larger expected absolute errors than any of the remaining six subtasks. In other words, on some tasks, large absolute bearing errors occur because of a systematic response bias (either to the left or the right), while, on other tasks, bearing errors tend to be less systematic and smaller in magnitude.

Log range errors depended significantly on training method ($F(3, 395) = 9.77$, $p < 0.001$), subtask ($F(12, 395) = 8.65$, $p < 0.001$), and subject ($F(31, 395) = 24.68$, $p < 0.001$). Scheffe post-hoc analysis indicated that the expected log range error in the VE condition (value of −0.022) was significantly larger than the expected log range errors in the Mod and RW conditions (values of −0.089 and −0.083, respectively). In the NVE task, the expected log range error was −0.033, a value that was not significantly different from any of the values in the other three conditions. In other words, across all tasks, subjects tended to underestimate the distance to the target. This tendency was less pronounced in the VE condition compared to the Mod or RW conditions. While this trend was significant, the magnitude of the difference was relatively small. ANOVA analysis also showed that subtask was a significant factor; however, once again, the magnitudes of the differences across subtasks are quite small. The expected log range values varied from −0.176 on subtask 11 to 0.016 on subtask 12. For most tasks, the expected log range errors were negative (subjects tended to underestimate the range). Only subtasks 5, 8, and 12 had positive expected values. Scheffe post-hoc analysis showed that the only significant differences among all pairwise comparisons arose when comparing subtask 11 with these subtasks.

An ANOVA on the absolute log range errors showed that training condition ($F(3, 395) = 18.69$, $p < 0.001$) and subject ($F(31, 395) = 6.07$, $p < 0.001$) were significant factors. However, subtask ($F(12, 395) = 1.75$, $p > 0.001$) was not a significant factor. Scheffe post-hoc analysis indicated that the absolute log range error in the RW condition (expected absolute log range error of 0.224) was significantly larger than the absolute log range error in each of the other three tasks (expected log range errors of 0.157, 0.140, and 0.145 for VE, NVE, and Mod conditions, respectively). Once again, although this effect is significant, the magnitude of the difference in absolute range errors between real world training and the other methods is not large.

Taken together, these results indicate that the differences in range errors across subtask are relatively unimportant. However, errors in the range judgments do depend on training method. Specifically, there is a general tendency to underestimate target range, but this tendency is least pronounced in the NVE condition. The magnitude of the log range errors is largest in the RW condition (compared to the other three conditions). Although robust and statistically significant, neither of these effects is large in an absolute sense.