Available online at www.sciencedirect.com

SCIENCE DIRECT®

Intelligence 33 (2005) 169–186

INTELLIGENCE

The relationships between cognitive ability and dynamic decision making

Cleotilde Gonzalez*, Rickey P. Thomas, Polina Vanyukov

Dynamic Decision Making Laboratory, Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, United States

Received 9 December 2003; received in revised form 7 July 2004; accepted 20 October 2004
Available online 28 December 2004

Abstract

This study investigated the relationships between cognitive ability (as assessed by the Raven Progressive Matrices Test [RPM] and the Visual-Span Test [VSPAN]) and individuals' performance in three dynamic decision making (DDM) tasks (i.e., regular Water Purification Plant [WPP], Team WPP, and Firechief). Participants interacted repeatedly with one of the three microworlds. Our results indicate a positive association between VSPAN and RPM scores and between each of those measures and performance in the three dynamic tasks. Practice had no effect on the correlation between RPM score and performance in any of the microworlds, but it led to an increased correlation between VSPAN score and performance in Team WPP. The pattern of associations between performance in microworlds and assessments of cognitive ability was consistent with the task requirements of the microworlds. These findings provide insight into the cognitive demands of dynamic decision making and the dynamics of the relationships between cognitive ability and performance with task practice.

Keywords: Cognitive ability; Dynamic decision making; Practice

1. Introduction

Dynamic decision making (DDM) shares many of the characteristics and complexities of real-world decision making. Like the decision making required in many real-world situations, DDM is

* Corresponding author. Tel.: +1 412 268 6242; fax: +1 412 268 6938.
E-mail address: conzalez@andrew.cmu.edu (C. Gonzalez).

0160-2896/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2004.10.002

characterized by multiple and interdependent real-time decisions that must be made in an environment that changes as a function of the sequence of actions, independently from exogenous events, or in both ways (Brehmer, 1992; Edwards, 1962). DDM and much of real-world decision making are dynamically complex because of the nonlinear relationships that exist among variables, multiple loops, and feedback delays (Sterman, 2000). Typical real-world examples of DDM include command and control in battle situations, firefighting, air traffic control, production scheduling, and emergency dispatch.

The study of individuals' differences is a traditional topic in DDM research (Brehmer & Dörner, 1993). Researchers working in this area both aim to characterize individual abilities that would help to explain performance and to explore the demands of DDM tasks. Although seemingly promising, this approach has yielded only limited success. Many studies have indicated that performance in dynamic tasks varies tremendously among individuals, but psychological assessments of cognitive abilities and personality have been unable to explain this variability (Brehmer, 1992; Rigas & Brehmer, 1999). This state of affairs in DDM research is surprising because many studies in psychology have documented greater correlation between performance and ability as task complexity increases (Ackerman, 1988; Kyllonen, 1985). On the basis of both this research in psychology and the complexity of DDM, one would expect greater correlation between DDM performance and general cognitive ability.

Rigas and Brehmer (1999) have proposed the *different-demands hypothesis* as an explanation for the weak correlations that researchers have observed between a measure of general fluid intelligence (Gf) (i.e., the Raven Progressive Matrices Test [RPM]) and DDM performance. This hypothesis suggests that dynamic tasks demand the performance of more complex mental processes than do intelligence tests. There is some laboratory support for the different-demands hypothesis. Joslyn and Hunt (1998) developed a task that could predict operators' performance in two real-world DDM domains: public safety dispatch (911 operators) and air traffic control (air traffic controllers), but no correlations existed between this task and an RPM-like measure of Gf (Joslyn & Hunt, 1998).

Recently, Rigas, Carling, and Brehmer (2002) reported significant correlations between RPM score and performance in two dynamic tasks (Rigas et al., 2002). These findings led Rigas et al. to offer the *low-reliability hypothesis* as a possible explanation of why prior research failed to establish an association between performance in DDM tasks and intelligence. The low-reliability hypothesis argues that intelligence scores have failed to correlate with performance measures in most dynamic tasks because the performance measures have suffered from low reliability. There is evidence that some past studies have suffered from poor reliability (e.g., α-coefficients <0.44) (as cited in Rigas et al., 2002). The reliability coefficients of the DDM performance measures reported in Rigas et al.—the only study to find a significant correlation between performance and intelligence in DDM tasks—were within acceptable limits, with all α-coefficients >0.77.

Both of the aforementioned hypotheses warrant further investigation. Low reliability is a concern for any study involving the performance of dynamic tasks (Funke, 1995) and, because very little is known about which cognitive abilities are necessary for successful performance of dynamic tasks, the different-demands hypothesis also appears well-founded. Our study, reported below, extends this previous research in two ways. First, we observed correlations between cognitive ability and performance in three DDM tasks involving realistic simulations in the laboratory setting (i.e., microworlds): Water Purification Plant (WPP) (Gonzalez, Lerch, & Lebiere, 2003), Firechief (Omodei & Wearing, 1995), and a team-oriented variation of WPP (Team WPP). We calculated the reliability of the performance measure used for each of these tasks. In addition to using the RPM measure of fluid intelligence, we used the Visual-Span Test (VSPAN), which measures working memory (WM) capacity, and identified any

correlations between individuals' performance in the microworlds and their scores on each of these two tests. We also investigated the possible effects of task practice on the cognitive ability and performance correlations. Most DDM researchers have failed to investigate changes in performance that are attributable to extended practice, despite reports suggesting that operators require significant amounts of practice to learn to control dynamic systems (Kerstholt & Raaijmakers, 1997). A heightened understanding of the dynamics of correlations between cognitive ability and performance may enable us to determine the information-processing demands of different dynamic tasks and, by so doing, to identify task differences and similarities.

2. Cognitive ability and DDM performance predictions

Evidence suggests that microworlds share several common characteristics, including *complexity*, *dynamics*, and *opaqueness* (Brehmer & Dörner, 1993). *Complexity* in this context refers to the number of variables that individuals must consider simultaneously while making a decision. *Dynamics* refers to the change in the decision state as a consequence of decisions and as an effect of environmental actions that are beyond the user's control. *Dynamics* also refers to the need for users to make decisions in real-time as the environment changes. *Opaqueness* refers to the hidden characteristics of a microworld that, although not explicitly presented to the operator, must be identified by users and incorporated into their decisions if they wish to perform well. *Dynamic complexity* is another characteristic of microworlds (Gonzalez, Vanyukov, & Martin, in press). As defined by Sterman (2000), *dynamic complexity* refers to the interrelations of system variables over time. These nonlinear relationships are created by the positive and negative feedback cause-and-effect loops that are characteristic of dynamic systems.

Although most microworlds embody these characteristics, research suggests that many microworlds differ from one another in terms of the degree to which they incorporate these attributes (Gonzalez et al., in press). For example, some microworlds involve a small number of variables but become very dynamically complex as the variables interact over time. Because microworlds differ in terms of their structural characteristics, successful performance in different microworlds is likely to require the use of different cognitive abilities (Gonzalez et al., in press). For the current study, we selected three microworlds that share the general characteristics described above and thus should elicit the use of similar cognitive abilities. The next section provides a detailed description of these microworlds, their similarities and differences, and their expected correlations with cognitive ability.

Correlation studies of cognitive ability have demonstrated several points that are relevant to the current study. First, researchers have identified clusters of cognitive tasks that define particular ability constructs (Snow, Kyllonen, & Marshalek, 1984). Specifically, these studies have shown that cognitive ability tests cluster by content, which typically is verbal, spatial, numeric or symbolic in nature. Furthermore, this research demonstrates that general abilities, such as intelligence, often encompass other more specific abilities, such as verbal, spatial, or numeric abilities (Snow et al., 1984). Most microworlds are graphical representations of actual systems. For example, MORO, a well-known microworld (as described in Strohschneider & Guss, 1999) depicts a small tribe of semi-nomads in the southern Sahara, and Lohhausen (as described in Funke, 1988) depicts a fictional city. The microworlds used in our study are also graphical and spatial in nature. For example, Firechief (Omodei & Wearing, 1995) simulates fire spreading through a landscape, and users must attempt to extinguish the fire as soon

as possible. On the basis of the prior research discussed above, we hypothesized that spatial cognitive abilities would correlate with performance in the microworlds used in our study.

Second, researchers have demonstrated that relationships between cognitive ability and performance strongly depend on the complexity of a task (Ackerman, 1988; Kyllonen, 1985). The complexity of a dynamic task is determined by the number of variables and the number and kind of relationships that exist among those variables (Brehmer & Dörner, 1993). Often, dynamic systems incorporate multiple variables that are closely related and affect each other via positive or negative feedback loops. Ackerman (1992) reported strong correlations between general cognitive ability and performance in a microworld called Air Traffic Control (ATC) (Ackerman, 1992). Wittmann and Süβ (1999) also reported significant correlations between cognitive abilities and personality measures and performance in complex simulations. In particular, Wittmann et al.'s work highlights the importance of working memory to performance in complex tasks. In light of the complexity of microworlds and these previously demonstrated results, we expected to observe a definite relationship between general cognitive ability and performance in the microworlds used in our study.

Third, the relationships between cognitive ability and performance appear to depend on the amount of task practice (Ackerman, 1988). In his theory regarding the ability-related determinants of skill acquisition, Ackerman posits that individuals pass through different phases of learning that are characterized by different cognitive demands. In essence, this theory predicts that initial performance demands the use of general and content abilities but, as a decision maker acquires additional skills needed to perform the task, the influence of these abilities on performance will diminish. Further investigations on this theory suggest that as task complexity increases the ability-performance correlations will be consistent across practice (Ackerman, 1992). On the basis of this theory and given the complexity of the tasks used in this study, we expected that general ability-performance relationships will be strong and consistent throughout task practice in all the microworlds.

3. Methods

3.1. DDM tasks: the microworlds

3.1.1. Water purification plant (WPP)

The 'Water Purification Plant' (WPP) is a computer-interactive simulation isomorph of a real-world scheduling task. The real-world task from which WPP has been developed is mail sorting in a large-scale organization, the United States Postal Service (USPS). WPP was developed with the parameters and variables involved in the USPS task (Lerch, Ballou, & Harter, 1997), but with a simplified interface that would help to quickly train participants. WPP has been described in detail in previous research and it has been used to study DDM and the cognitive processes involved in making decisions in dynamic environments (Gonzalez et al., 2003). Here we summarize the task.

Fig. 1 shows the layout of WPP. In WPP, the decision maker plays the role of a Water Purification Plant operator, whose goal is to distribute water to different locations on time. The amount of water to distribute is defined by a scenario that describes the pattern of water arrival to the system. The scenario defines the arrival time, the amount of water dispersed, and the destination tank. For example, an entry in the WPP scenario may indicate that at 2:02 PM 10 gal of water will be assigned to tank 2. The scenario is unknown to

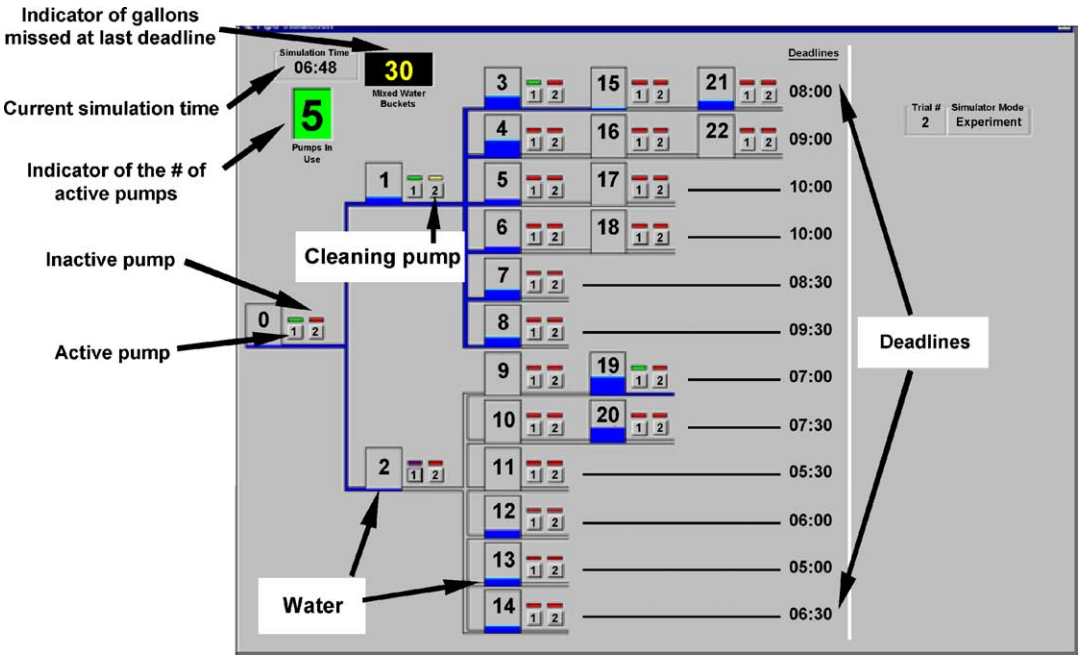

Fig. 1. WPP (individual version).

the participants. Performance in WPP is measured by the cumulative number of gallons of water that missed the deadlines.

Operators open or close pumps assigned to each tank to let the water flow to the following tank in the chain that leads to a deadline (water flows only from left to right). There are 23 tanks in the system interconnected in a tree structure and two pumps per tank, but only five pumps can be active at the same time in the system (this is the total number of sorting machines available in the USPS office, participants are told that electricity constraints prevent them from using more than five pumps concurrently). Consecutive tanks such as tank 3, tank 15, and tank 21 make a chain and thus have the same deadline, 8:00 PM.[1] The simulation provides an indicator to track the number of pumps in use (shown at the top right corner in Fig. 1). Each of the pumps delivers water at a rate of 1 gal every two simulation minutes; however, when two pumps in one tank are active the delivery rate is 1 gal/min. Each pump may be on one of four statuses indicated by different colors on the screen: off (red), on (green), cleaning (yellow) or in queue (purple). Pumps turn to cleaning status after they are turned off (either by the operator by clicking on an active pump or by the system when there is no more water to pump in a tank). Each pump takes 10 min of simulation time to finish cleaning. While pumps are cleaning, other pumps (within the five-pump limit) may be queued to start as soon as a cleaning pump is available.

The main performance indicator in this task is the total number of gallons of water remaining in the system upon expiration of the deadlines (i.e., the gallons of water missed). After each deadline, the simulation shows the total number of gallons of water missed for that deadline (see top left corner of Fig. 1). The missed gallons score is updated after each deadline up to and including the final deadline of

[1] We added numbers to the tanks to facilitate our description of the picture in the text. These numbers were not visible to the study participants.

the simulation (10:00 PM), at which time the total number of missed gallons is shown on the screen. The best performance in WPP is 0, which indicates that the operator removed all of the water before expiration of the deadlines.

3.1.2. Team WPP

Team WPP extends the individual version of WPP so as to involve two operators (a host and a client), who must work together to meet the deadlines while using limited resources (each of the players can activate only five pumps at any given time). Team WPP allows the two members of a team to cooperate by transferring water to each other and to communicate by using a chat tool to write messages (see Fig. 2). While operating Team WPP, each team member can view in real time a constantly updated version of the other team member's screen. The left side of the screen shows the team member's own simulation, while the right side shows his or her partner's simulation. The goal in Team WPP is the same as that in the individual version of WPP: pump water through the tanks before expiration of the deadlines, which appear on the right side of the layout. Both team members have identical deadlines and must pump the water through identically structured systems; neither team member can operate more than five pumps concurrently. To increase the need for collaboration, the preset scenario in Team WPP distributes the water unevenly between the two members, giving a small initial amount of water to one

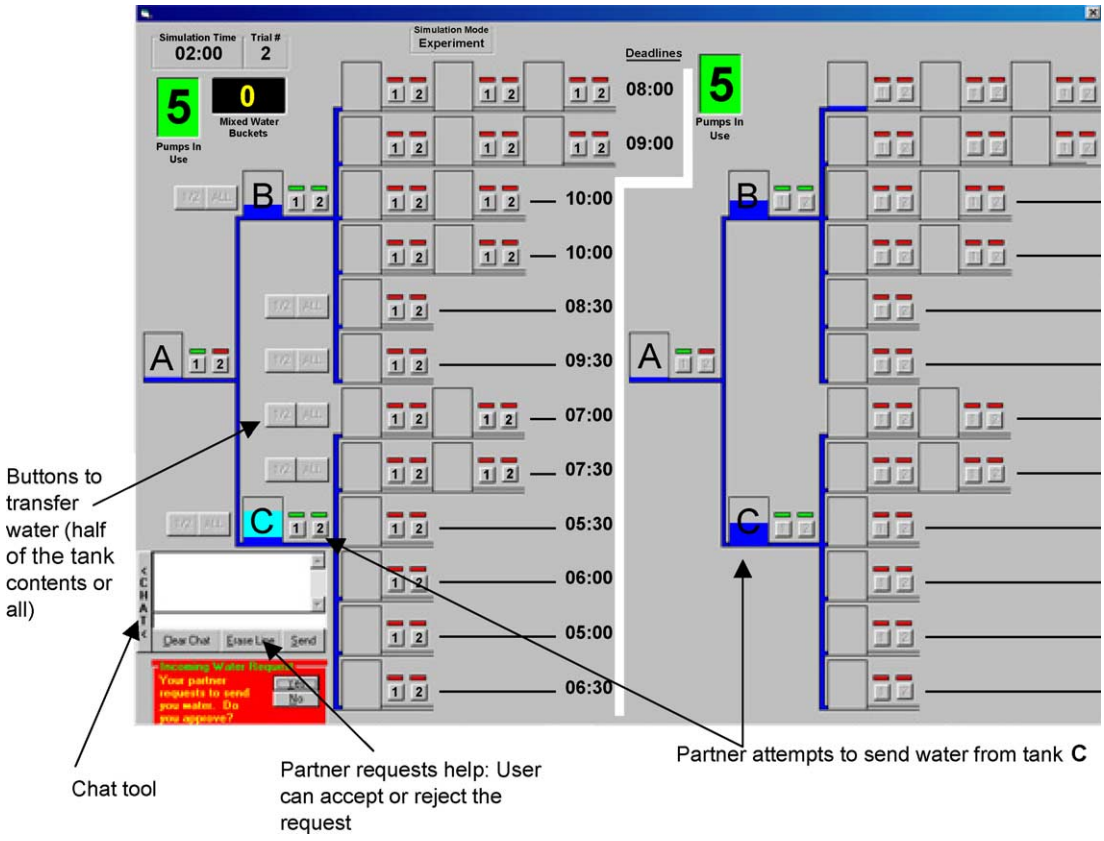

Fig. 2. WPP (team version).

member and a large amount to the other. Team members can assist one another by using buttons associated with certain tanks to send and receive water from one system to the other (see Fig. 2). For example, the host can click on one of the buttons to send either the full amount of water contained in a tank or half of that amount to the client. The client receives a warning indicating that the host is sending water, and the client can elect to view the amount of water before accepting or rejecting the transfer. The host and the client can send water back and forth to each other only from the same pre-assigned tanks. The measure of performance in Team WPP is the total gallons of water missed by both members of the team. There is no indication of individual performance.

3.1.3. Firechief

The layout of Firechief, a microworld developed by Omodei and Wearing (1995), is shown in Fig. 3. This microworld simulates fire events that spread continuously to adjacent unburned areas at a rate that depends on factors such as the wind conditions and the type of landscape. Users try to extinguish the fire as soon as possible. To do so, they must use their limited resources (helicopters and trucks) wisely. To parallel the limited resources in WPP and Team WPP, we assigned five units (three helicopters and two trucks) in the Firechief scenario used for this study. By using controlled mouse movements and key strikes, Firechief users can move units to a specified landscape segment and drop water on that segment or replenish their water supply.

Users can move helicopters (yellow squares) or fire trucks (white squares) toward the area of fire outbreak. Upon their arrival, these vehicles begin to extinguish the fire. At any one time, different vehicles may be inactive (because their water supply is depleted), in motion toward a location, in the process of extinguishing the fire, or refilling with water. The need for users to view and remain mindful of additional information while controlling the status of different units complicates the task and can hinder performance. Wind direction, as shown in the top left corner of the screen, influences the direction in which the fire spreads. The type of landscape, distinguishable (based on color) as forest, grasslands, clearings, or housing areas, also influences the strength of the fire and the rate at which it spreads. Participants must realize that vehicles deplete a resource (water) when used to extinguish the fire. After a vehicle runs out of water, it becomes inactive and the user must refill it. The refilling process results in a delay during which the vehicle cannot take part in the fire-extinguishing process. Firechief is presented in real time, and performance is measured as the percentage of unburned landscape (burned areas are marked in black) at the end of the simulation.

3.1.4. Similarities and differences between microworlds

WPP and Firechief are similar in many respects. First, both tasks are dynamic resource allocation tasks. In both systems, exogenous events partially define the status of the system (water in WPP and fire in Firechief). Users' actions, which are restricted by limited resources, also affect the status of the system. Second, both tasks are complex. WPP and Firechief involve multiple variables (e.g., pumps, water, deadlines; wind direction, fire, locations), and many of the relationships among these variables are nonlinear. In Firechief, for example, the spread of the fire—as dictated by the flammability of the landscape and the wind direction—may influence the optimal allocation of the units. Third, both tasks are opaque in the sense that many characteristics of the systems are not visually apparent, but must be identified by user inference. In WPP, for example, users must estimate the time needed to pump water out of a tank; users often require practice to learn how to accurately estimate this factor.

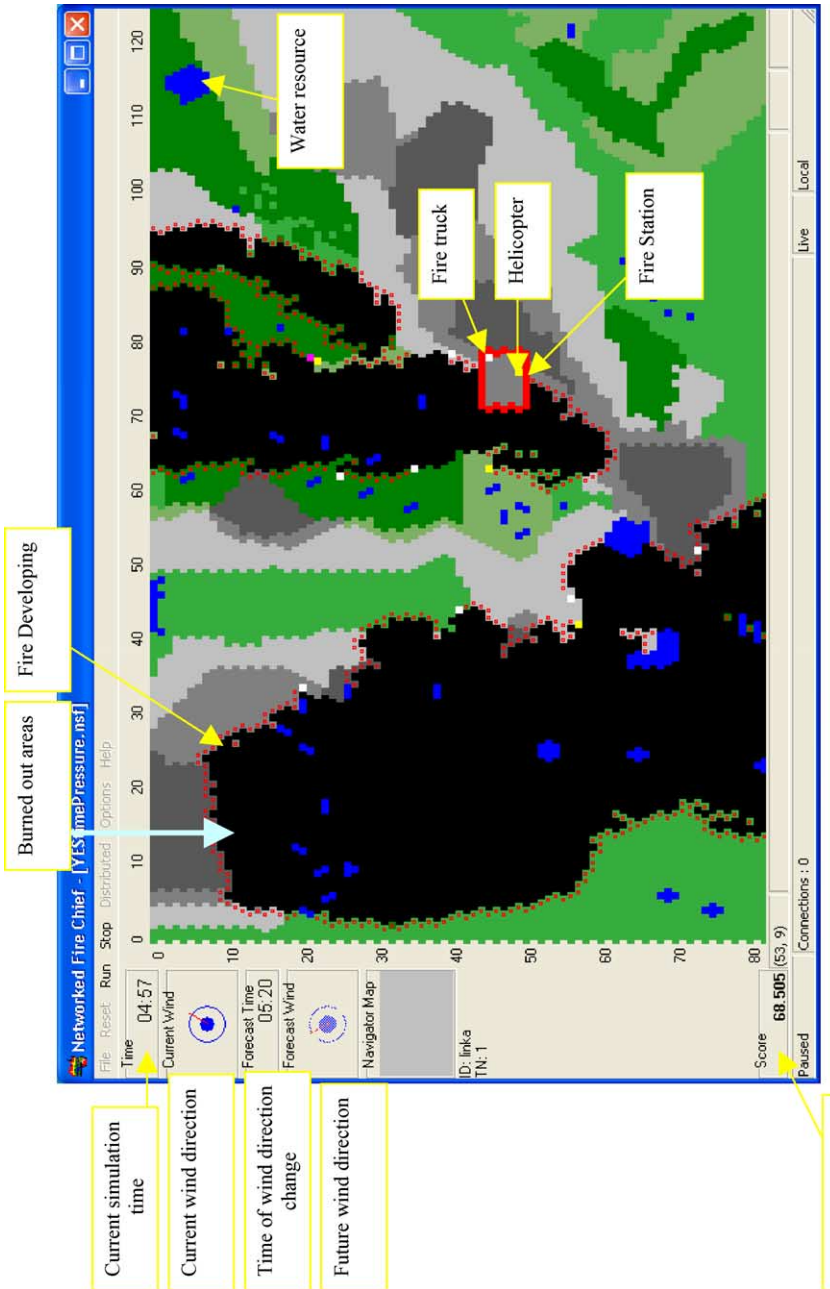

Fig. 3. Firechief.

Although it is relatively easy to identify the similarities between WPP and Firechief, it is difficult to see the differences; yet it is the differences in the tasks' structures that may have the greatest effect on a user's performance and on the relationship between performance and cognitive ability. First, it is obvious that the individual microworlds are different from the Team WPP task. Team WPP is more complex than individual WPP because, in addition to the requirements of individual WPP, Team WPP demands team coordination and visual management not only of the status of one's own task but also of the status of a partner's task. For this reason, we expected stronger correlations between cognitive ability and performance in Team WPP than in individual WPP or Firechief.

The differences between individual WPP and Firechief are less obvious. By analyzing the tasks, we have determined that WPP has a more static structure than Firechief. Neither the tree-like layout of WPP nor the deadlines vary during a simulation run. The deadlines in WPP encourage a certain sequence of pump activation and deactivation. The structure of WPP provides a user with cues (i.e., deadlines and chain values) that the user may use to identify the subgoals and rank them in terms of importance. In addition, the cleaning delays in WPP are a fixed length (10 s), whereas the delays in Firechief depend on how much time a user takes to realize that a vehicle is inactive, how far the user must send the vehicle to obtain more water, and how far the vehicle must travel to return to the fire itself. Firechief users must always be aware of the status of the different vehicles to avoid wasting too much time due to idle equipment. These time- and distance-dependent delays force Firechief users to coordinate their decisions by remembering a large amount of information needed to organize the subgoals of the simulation efficiently (e.g., which vehicle will require maintenance next). Because of these differences in task structures, we expected WPP users to depend less than Firechief users on WM capacity for coordination functions (e.g., the storage of subgoals) and for acquisition and storage of partial solutions relevant to solving the task (i.e., developing a strategy for meeting all the deadlines). Thus, we expected to observe stronger correlation between WM capacity and performance in Firechief than in WPP.

3.2. Cognitive ability tests

We selected two measures of cognitive ability: the RPM (standard or advanced version) (Raven, 1962, 1977), a measure of fluid intelligence, and the VSPAN, a measure of visual WM (Shah & Miyake, 1996). RPM was selected because of its inability to explain the variability in performance of DDM tasks in previous studies (Rigas & Brehmer, 1999), and because according to psychological research this measure is expected to correlate highly with performance in complex task (Ackerman, 1988). VSPAN was selected because the tasks demand WM capacity and in particular visual ability, due to their graphical representations.

Fluid intelligence influences individuals' ability to deal with novelty and to adapt their thinking to solve new cognitive problems (Carpenter, Just, & Shell, 1990). Researchers have analyzed the cognitive processes underlying fluid intelligence (as measured by RPM) to determine processing characteristics shared by individuals who score high on the RPM and processing characteristics shared by individuals who score low on this test (Carpenter et al., 1990). The processes that distinguish these two sets of individuals are the ability to induce abstract relations and the ability to use WM to dynamically manage a large set of problem-solving goals (Carpenter et al., 1990). In light of the characteristics of DDM tasks, we hypothesized that fluid intelligence would largely determine performance in the microworlds used in our study.

The RPM test consists of visual analogy problems. Each problem presents a 3×3 matrix in which one of the entries (out of nine) is missing. Test takers must select the entry that best completes the matrix from a set of eight choices at the bottom of the page. The test contains five sets of 12 questions each for a total of 60 questions, which appear according to their degree of difficulty (more difficult questions appear at the end of the test). The total number of correct answers serves as a measure of the test taker's analytical ability.

WM is the construct underlying many functions that are likely to influence performance in dynamic complex tasks. Süβ, Oberauer, Wittmann, Wilhelm, and Schulze (2002) argue that WM is responsible for three primary functions. First, WM influences individuals' ability to simultaneously store and process information (Daneman & Carpenter, 1980). Second, WM is required for the performance of supervisory functions, such as monitoring mental operations, controlling their efficiency, activating task-appropriate schema, and inhibiting task-inappropriate schema (Shah & Miyake, 1996; Turner & Engle, 1989). Third, WM plays a role in the integration and sequencing of mental operations and the prioritization of subgoals over extended periods of time. This ability is likely to affect performance in complex dynamic tasks that require both temporally extended coordination of multiple steps of processing and the storage of intermediate products of computation and subgoals (Verguts & de Boeck, 2001).

VSPAN measures an individual's ability to simultaneously store and process visual or spatial information. Individuals are presented with sets of letters and are asked to perform a spatial transformation (i.e., mental rotation) of the letters while simultaneously keeping track of spatial information (i.e., the top of the letter). Test takers must try at the end of the set to recall the orientation of all letters in the set. The test contains a total of 70 letters: 5 two-letter sets, 5 three-letter sets, 5 four-letter sets, and 5 five-letter sets. To score this task, we used a computer program, as detailed by Shah and Miyake (1996), to define each participant's visual span in terms of the total number of correctly recalled letters (70 maximum) (Shah & Miyake, 1996). Users required approximately 15 min to complete the VSPAN (including instructions).

3.3. Participants

Participants were recruited from local universities using electronic communication boards. Individuals gave signed, informed consent to participate voluntarily in this study, approved by the institution's review board. Their participation involved repeated interaction during 3 days (2 h/day) with one of the three microworlds: WPP, Team WPP, and Firechief. They were paid $50 after their participation. All participants completed the RPM (standard or advanced versions) and the VSPAN tests.[2] For Team WPP, we determined the cognitive ability measure by taking the average of the team members' scores on each test. This so-called *additive form of aggregation* is appropriate when team members can compensate for one another in team performance (LePine, Hollenbeck, Ilgen, & Hedlund, 1997).[3]

The total number of WPP participants used for our analysis was 74 (39 males and 35 females; mean age=21.70 years), the total number of WPP teams was 28 (56 total participants; 34 males and 22 females; mean age=24.05), and the total number of Firechief participants was 16.[4] The sample sizes in

[2] Firechief participants received an advanced version of RPM. We converted their scores to correspond to scores on the standard version of the test by using the methods described in Raven (1977).

[3] We also calculated several other aggregation measures (e.g., minimum member score, maximum member score, host score, client score, etc.). We found strong correlation between all the aggregate measures (>0.6) and between each aggregate measure and the additive aggregation measure.

[4] We did not collect demographic data from Firechief participants.

this study are low, in particular for Firechief, compared to other individual differences studies. However, it will be shown below that this sample size is sufficient to detect the relationships between cognitive abilities and performance in this microworld.

3.4. Procedure

During the first day, participants in the individual and Team WPP experiments completed the VSPAN and the RPM. All WPP participants then underwent WPP training. Participants received both instruction on how to operate the simulation interface and information about the task goals. During the instruction period, participants received information regarding deadlines, simulation time, and the paths of water travel. Participants learned that different amounts of water could enter any of the tanks at any time from outside the system, but they received no information regarding the amount of water to process or the time of water arrival. Participants were told to do their best to process all the water that appeared within the system but were not given any particular solution strategy. Additionally, the Team WPP participants were taught how to communicate with and transfer water to their partners. All participants (WPP and Team WPP) completed a practice trial during which we ran the simulation at the slowest possible pace to give them time to ask questions and experience all parts of the simulation. In Team WPP, participants were allowed to use the chat tool to communicate with their partners and to send and receive water from their partners via the send buttons located next to certain tanks. They were not allowed to communicate with each other in any other way. After the practice session, participants in both sets of experiments began to perform complete WPP trials. Participants in individual WPP performed the task 10 times, whereas participants in Team WPP performed the task 12 times.

Firechief participants completed the VSPAN and the RPM during the first day of the experiment. They next completed a short practice session of Firechief (approximately 7 min in length). Like the WPP practice session, the Firechief practice session was run at a slow pace. Participants received detailed instructions describing the landscape, the simulation, and how to use the vehicles, but they were not given any strategies by which to extinguish the fire effectively. After completing the practice session, the participants began to perform complete Firechief trials. They completed a total of 16 trials.

4. Results

We conducted several analyses to investigate how accurately the explanatory variables RPM score and VSPAN score predicted performance in the three tasks (i.e., individual WPP, Team WPP, and Firechief). First, we used a variety of methods to assess the normality of the data (i.e., bivariate scatterplots, skewness statistics, kurtosis statistics, quintile–quintile plots, normal-probability plots, and Shapiro-Wilk tests of normality). Although multiple regression is robust to moderate violations of normality and moderate violations of homoscedasticity (Tabachnick & Fidell, 2001), we observed substantial negative skew in the RPM score and individual WPP performance distributions, which indicates marked departures from normality. Using the individual WPP data, we performed a square-root transformation on the reflected RPM score and WPP performance distributions, and it successfully attenuated the negative skew and brought the data in line with normality assumptions. The diagnostics revealed no substantial deviations from normality in the Team WPP data and only mild deviations from normality in the Firechief data, which did not demand a transformation. Second, we examined all the

Table 1
Descriptive statistics

Task	*M*	S.D.	*N*	Skewness		Kurtosis		Shapiro-Wilk	α
				Statistic	S.E.	Statistic	S.E.		
Individual WPP									
VSPAN	37.5	13.11	74	−0.111	0.279	−0.394	0.552	0.993	0.88
RPM	2.54	0.73	74	0.399	0.279	0.480	0.552	0.973	–
(Raw score)	(54.0)	(4.0)		(−1.300)		(2.23)		(0.903**	–
Performance	7.81	2.78	74	0.147	0.279	−0.148	0.552	0.984	0.96
(Raw score)	(946)	(45.6)		(−0.976)		(0.474)		(0.919**)	(0.960)
Team WPP									
VSPAN	34.1	9.86	28	0.926	0.441	−0.224	0.858	0.968	0.94
RPM	27.0	1.59	28	−0.730	0.441	0.524	0.858	0.948	0.94
Performance	864	47.5	28	0.202	0.441	012	0.858	0.981	0.96
Individual Firechief									
VSPAN	31.9	14.5	15	−0.287	0.580	−1.02	1.12	0.945	0.93
RPM	25.8	5.84	15	−0.492	0.580	0.519	1.12	0.976	0.59
Performance	84.9	7.74	15	−1.05	0.580	0.778	1.12	0.914	0.97

Data in parentheses represent the statistics for the untransformed score distributions.
** $p \leq 0.01$.

data sets for outliers by using preliminary regression based on calculations of the Mahalanobis distance for each participant. This method identifies units with Mahalanobis distance values greater than 16.27, chi square (χ^2) critical value for p=0.001, and df=3 as outliers (Mertler & Vannatta, 2002). The outlier analysis revealed no suspect data points.

Table 1 presents the descriptive statistics for VSPAN score, RPM score, and performance in each of the three microworlds. The RPM scores and standard deviations are within expected ranges, as presented in the relevant literature (Engle, Tuholski, Laughlin, & Conway, 1999), and the average RPM scores of the participants are within the 60th percentile for 18–22-year-old adults in the United States (Raven, Raven, & Court, 1993). The VSPAN scores and standard deviations also are within the expected ranges, as outlined in the literature (Shah & Miyake, 1996). In addition, the high α-coefficients for all variables support the reliability of the tests and performance measures used in the three tasks.[5]

Table 2 reports the Pearson correlations between VSPAN score, RPM score, and average performance in each of the three tasks. All correlations are statistically significant. VSPAN score correlates significantly with RPM score, and both scores correlate significantly with average performance in each of the tasks (i.e., operators that scored higher on the VSPAN and RPM performed better in the tasks). The correlation effects are moderate (r=~0.3) for individual WPP, and large (r>0.5) for Team WPP and Firechief. The effect of VSPAN score and that of RPM score are very similar in the WPP task (r=~0.3 in both tests for individual WPP and r=~0.6 in both tests for Team WPP), but they differ in Firechief (r for VSPAN=0.82, whereas r for RPM=0.61).

[5] We could not calculate reliability coefficients for the RPM scores in the individual WPP group because only the summary data for this study are available. However, RPM score correlates significantly with both VSPAN and WPP performance. Thus, it is reasonable to assume that the reliability is high enough for the measure to correlate significantly with other measures. Moreover, prior research suggests that RPM scores are generally reliable and valid (Raven, 1977).

Table 2
Correlation analyses

Task	RPM	Performance
Individual WPP		
VSPAN	0.352*	0.333*
RPM		0.331*
Team WPP		
VSPAN	0.647*	0.637*
RPM		0.618*
Individual Firechief		
VSPAN	0.702*	0.820*
RPM		0.605*

* $p \leq 0.05$.

We used standard linear regression models to determine the contribution of each of the cognitive tests (i.e., RPM and VSPAN) to performance in each of the tasks. We analyzed each task with three models: one with VSPAN, one with RPM, and one with both measures. Table 3 summarizes the models. Statistical evaluation of tolerance and variance inflation factor (VIF) indicated no substantial multicollinearity. The results generated by all regression models were significant ($p<0.05$). Our findings demonstrate that each of the cognitive test scores can independently predict performance in each of the three tasks. However, use of the model based on both predictors reveals that, particularly in Team WPP and Firechief, RPM score accounts for only a negligible amount of the performance variance unaccounted for by VSPAN score. For instance, the R^2 value for the regression model predicting Firechief performance based solely on VSPAN score increased only by 0.01 after incorporation of RPM score within the model. These results indicate that individual WPP performance is associated with the

Table 3
Summary of regression models

Task	B_0	$t(B_0)$	B_1	$t(B_1)$	B_2	$t(B_2)$	R^2	F	VIF	pr
Individual WPP										
VSPAN (n=72, df=1)	10.4	11.2**	0.070	3.00**	–	–	0.110	8.98**	–	0.245
RPM (n=72, df=1)	4.63	4.17**	–	–	1.25	2.98**	0.110	8.88**	–	0.243
FULL (n=71, df=2)	7.41	4.36**	0.052	2.13*	0.925	2.11*	0.160	6.92**	1.141	
Team WPP										
VSPAN (n=26, df=1)	759	29.4**	3.07	4.22**	–	–	0.410	17.8**	–	0.396
RPM (n=26, df=1)	367	2.95**	–	–	18.5	4.01**	0.380	16.1**	–	0.350
FULL (n=25, df=2)	512	3.81**	1.97	2.16*	10.6	1.87	0.480	11.5**	1.72	
Individual Firechief										
VSPAN (n=13, df=1)	70.9	24.1**	0.440	5.17**	–	–	0.67	26.8**	–	0.560
RPM (n=13, df=1)	64.2	8.31**	–	–	0.800	2.74*	0.37	7.52*	–	0.040
FULL (n=12, df=2)	69.6	11.7**	0.416	3.37**	0.078	0.253	0.68	12.5**	1.97	

* $p \leq 0.05$.
** $p \leq 0.01$.

Table 4
Tests of linear trends

Task	R^2	R^2_{Adj}	t
Individual WPP			
VSPAN	0.01	−0.12	−0.25
RPM	0.34	0.25	4.03
Team WPP			
VSPAN	0.40	0.34	2.57*
RPM	0.17	0.09	1.44
Individual Firechief			
VSPAN	0.59	0.56	−4.48*
RPM	0.15	0.09	−1.55

* $p \leq 0.05$.

cognitive abilities tested by both the VSPAN and RPM, whereas Team WPP and Firechief performance is associated primarily with the cognitive abilities tested by the VSPAN.

4.1. Changes in ability and performance correlations with practice

To assess the ability of VSPAN score and RPM score to predict performance by individuals given the opportunity to practice a task, we calculated the correlation between each of the measures and performance in each trial of each task. All correlations were significant ($p<0.05$). We also used the slope of the regression line to identify linear trends of these correlations. This is a common analytical technique in skill-acquisition research (Ackerman, 1992). The results of this analysis are shown in Table 4. These results indicate that as expected, the correlations between performance and RPM scores for all three tasks remain stable regardless of task practice ($p>0.05$). Also, although the correlations between VSPAN score and performance in individual WPP and Firechief remained stable over time (i.e., with practice), the correlations between VSPAN score and performance in Team WPP became higher over time. Fig. 4 depicts these correlations between performance per trial and RPM and VSPAN scores for the three microworlds.

5. Discussion

Like the findings reported by Rigas et al. (2002), our results refute the strongest formulation of the different-demands hypothesis, which predicts close to zero correlation between intelligence test scores and performance in microworlds (Rigas et al., 2002). The reliability coefficients for the performance measures in WPP and Firechief were all very high (>0.94). Furthermore, our analysis revealed significant correlations between these performance measures and cognitive ability (as tested by the VSPAN and the RPM), findings that support the low-reliability hypothesis. However, our results also extend current understandings of the reliability and validity of performance measures in microworlds.

This research provides insight into the cognitive demands of DDM and the dynamics of the cognitive ability and performance correlations in DDM tasks. The significant correlations between cognitive

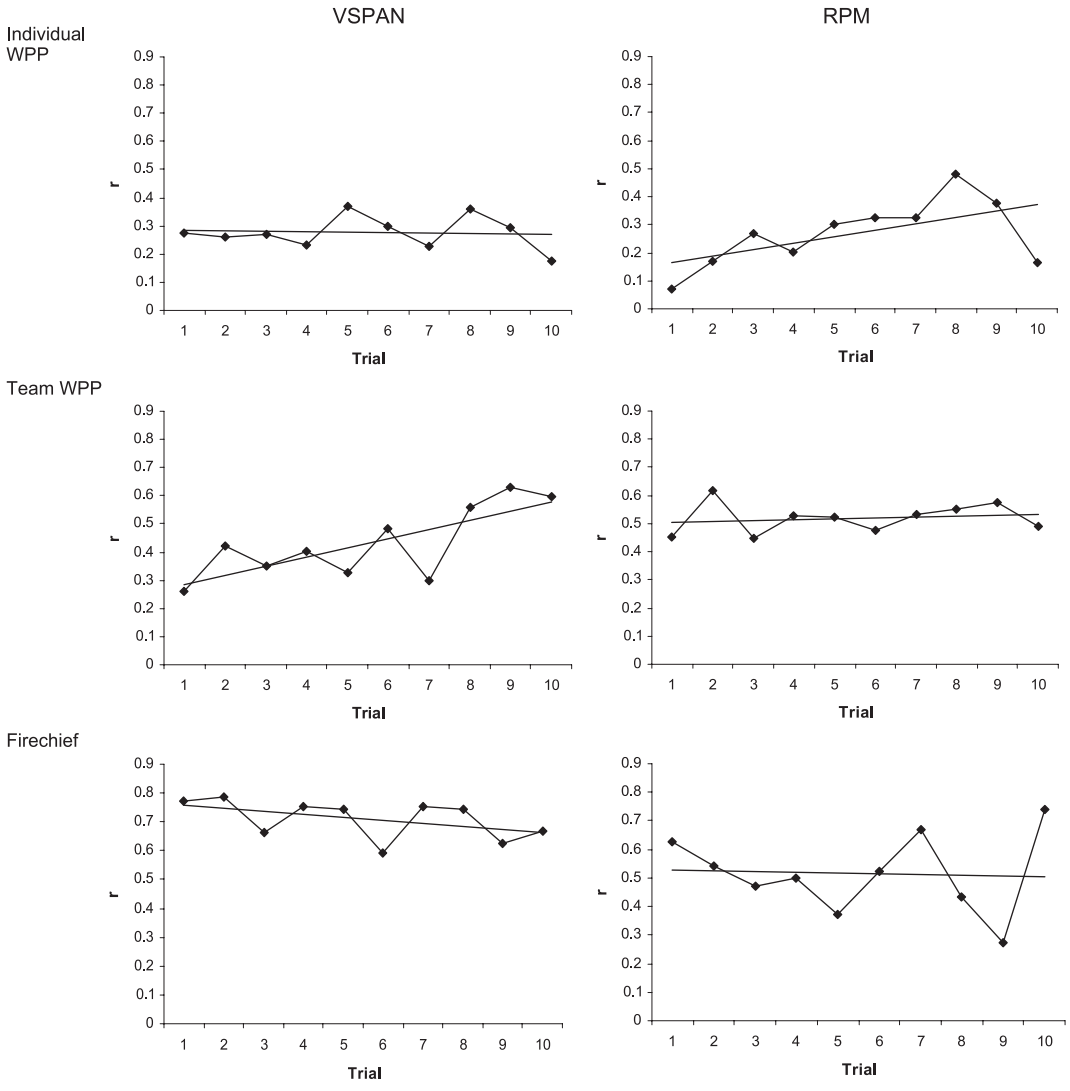

Fig. 4. Correlations between cognitive ability (as measured by VSPAN and RPM) and performance in the three microworlds in each trial and with practice.

ability and performance in all microworlds reflect certain shared characteristics of these tasks: complexity, dynamics, and opaqueness (Brehmer & Dörner, 1993). However, our results also suggest differences among these tasks.

The correlations between performance in individual WPP and cognitive test scores (VSPAN and RPM) remained consistent regardless of task practice; however, the correlations between performance in Firechief and cognitive test scores were less consistent. VSPAN score predicted more strongly Firechief performance than did RPM score, and the difference between the predictive capacity of VSPAN score and that of RPM score was higher in Firechief than in any of the other microworlds. Regression results indicate that VSPAN score is a more reliable predictor of performance than is RPM score in general, but

this difference was less noticeable in individual WPP than in Firechief. These results are consistent with the literature. For example, Schoppeck (1991) reported a correlation between spatiotemporal ability and performance in FEUER, a firefighting microworld, similar to the correlation reported here in Firechief. Ackerman (1992) also demonstrated strong and sustained correlations between spatial ability tests and performance in the ATC microworld. Careful evaluation of the ATC task used by Ackerman reveals many similarities between ATC and Firechief. Both tasks are spatial, and both require users to exhibit some perceptual speed to keep track of the rapid changes occurring on the screen and some perceptual psychomotor abilities to react quickly to these rapid changes.

We attribute the different results in the three microworlds used in this study to variances in the task characteristics. As predicted, performance in the WPP task appeared to demand less WM capacity than did performance in the Firechief task because the structure of the former task is more static and provides more cues regarding how to organize subgoals. Because Firechief requires more high-speed coordination and a greater ability to deal with a lack of task structure, unlimited response options, time pressure, etc., it caused the users to rely heavily on their WM.

Our study also revealed that task practice had microworld-specific effects on the correlations between cognitive ability and task performance. The influence of fluid intelligence on performance remained constant over time in all three microworlds. The correlation between performance and RPM score was stronger in Firechief than in the other two tasks, suggesting more complex processing requirements; however, the correlation did not vary with practice in any of the three. This result also supports Ackerman's (1992) hypothesis, which states that tests of general cognitive ability should accurately predict performance (regardless of task practice) in complex tasks that require a high degree of dynamic information processing.

Correlations between VSPAN score and performance in individual WPP and Firechief remained consistent over time, but these correlations *increased* with increasing practice in Team WPP. Thus more practice led users of Team WPP to increasingly depend upon the cognitive abilities tested by the VSPAN. This finding can be explained by an analysis of the demands of team microworlds versus those of individual microworlds. Teams in our study consisted of a pair of co-located, nonhierarchically organized, homogenous operators. Team WPP, as discussed in the analyses of similarities and differences among tasks, requires a user to keep track of continuous changes in his or her simulation while concurrently assessing continuous changes in the simulation of his or her partner. Decisions made by a user are necessarily influenced by the status of the individual's own task as well as that of a partner's task. Thus, the team WPP task demands more visual management than do the individual tasks. The results from our study are consistent with a large number of findings that show positive relationships between general measures of cognitive ability and both team performance and team effectiveness (LePine et al., 1997; Neuman & Wright, 1999).

In summary, the failure of most studies to demonstrate significant correlations between performance in microworlds and measures of cognitive ability has led many researchers to question if the cognitive abilities assessed by the tests are the same as those required for good decision making in microworlds. Researchers also have questioned the reliability and validity of microworld performance measures. Our results demonstrate that it is possible to identify reliable measures of microworld performance and that the skills required to complete some cognitive ability tests parallel those necessary for good performance in certain microworlds. Finally, our study indicates that dynamic tasks that are very similar in terms of their complexity, dynamics, and opaqueness may still place different cognitive demands on users; this results in differing degrees of correlation between cognitive ability and performance in the three tasks. A

theoretical structure of abilities, as described in Snow et al. (1984) and in the skill acquisition theory proposed by Ackerman (1988), may enable DDM researchers to select reliable measures that correspond to the abilities needed for good performance in DDM tasks.

Acknowledgements

This research was supported by the Multidisciplinary University Research Initiative Program (MURI; N00014-01-1-0677) and by the Advanced Decision Architectures Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory (DAAD19-01-2-0009). We thank Earl Hunt, Roberto Colom, Douglas Detterman, Werner Wittman, Phillip Ackerman, and an anonymous reviewer for their comments on previous versions of this paper.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology. General*, *117*(3), 288–318.

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinant. *Journal of Applied Psychology*, *77*(5), 598–614.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, *81*(3), 211–241.

Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior*, *9*(2–3), 171–184.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, *97*(3), 404–431.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors*, *4*, 59–73.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology. General*, *128*, 309–331.

Funke, J. (1988). Using simulation to study complex problem solving. *Simulation & Games*, *19*(3), 277–303.

Funke, J. (1995). Experimental research on complex problem solving. In P. Frensch, & J. Funke (Eds.), *Complex problem solving: The european perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Gonzalez, C., Vanyukov, P., & Martin, M. (in press). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*.

Joslyn, S., & Hunt, E. (1998). Evaluating individual differences in response to time–pressure situations. *Journal of Experimental Psychology. Applied*, *4*, 16–43.

Kerstholt, J. H., & Raaijmakers, J. G. W. (1997). Decision making in dynamic task environments. In R. Ranyard, R. W. Crozier, & O. Svenson (Eds.), *Decision making: Cognitive models and explanations*. Norwood, NJ: Ablex.

Kyllonen, P. C. (1985). *Dimensions of information speed* (No. AFHRL-TP-8-56). Brooks Air Force Base, TX: Air Force Systems Comman.

LePine, J. A., Hollenbeck, J. R., Ilgen, D. R., & Hedlund, J. (1997). Effects of individual differences on the performance of hierarchical decision-making teams: Much more than g. *Journal of Applied Psychology*, *82*, 803–811.

Lerch, F. J., Ballou, D. B., & Harter, D. E. (1997). Using simulation-based experiments for software requirements engineering. *Annals of Software Engineering*, *3*, 345–366.

Mertler, C., & Vannatta, R. (2002). *Advanced and multivariate statistical methods*, 2nd ed. Los Angeles: Pyrszak Publishing.

Neuman, G. A., & Wright, J. (1999). Team effectiveness: Beyond skills and cognitive ability. *Journal of Applied Psychology*, *84*, 376–389.

Omodei, M. M., & Wearing, A. J. (1995). The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods, Instruments, and Computers*, *27*, 303–316.

Raven, J. (1962). *Raven standard progressive matrices test*. Oxford: Oxford Psychologists Press.

Raven, J. (1977). *Advanced raven progressive matrices*. Oxford: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (1993). *Advanced progressive matrices: Section 1*. Oxford, England: Oxford Psychologists Press.

Rigas, G., & Brehmer, B. (1999). *Mental processes in intelligence tests and dynamics decision making tasks*. London: Lawrence Erlbaum Associates.

Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, *30*(5), 463–480.

Schoppeck, W. (1991). Spiel und wirklichkeit-reliabilitat und validitat von verhaltensmustern in komplexen situationen. *Sprache & Kognition*, *10*, 15–27.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology*, *125*, 4–27.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Eds.). *Advances in the Psychology of Human Intelligence*, *vol. 2* (pp. 47–103). Hillsdale, NJ: Erlbaum.

Sterman, J. (2000). Learning in and about complex systems. *Reflections: The SoL Journal*, *1*(3), 24–51.

Strohschneider, S., & Guss, D. (1999). The fate of the MOROS: A cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology*, *34*(4), 235–252.

Süβ, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, W., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*, 261–288.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*, 4th ed. Boston: Allyn and Bacon.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154.

Verguts, T., & de Boeck, P. (2001). On the correlation between working memory capacity and performance on intelligence tests. *Learning and Individual Differences*, *13*(1), 37–56.

Wittmann, W. W., & Süβ, H. M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem solving performance via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants*. American Psychological Association.