Many current theories emphasize the importance of team mental models1 to team training and performance (e.g., Mohammed & Dunville, 2001). Mental models are based on the premise that people organize information into patterns that reflect existing relationships between concepts and the features that define them (Johnson-Laird, 1983). As such, the measurement of mental models goes beyond the amount of declarative knowledge acquired but instead refers to an organized understanding or mental representation of that knowledge (Cannon-Bowers, Salas, & Converse, 1993; Klimoski & Mohammed, 1994). It has been hypothesized that team mental models enable team members to form common expectations, coordinate actions, adapt their behaviors to task demands, facilitate information processing, provide support, and diagnose deficiencies. As such, team mental models are an emergent characteristic of teams that influences both team processes (e.g., communication, conflict) and team outputs (e.g., performance; Klimoski & Mohammed, 1994; Kraiger & Wenzel, 1997; Marks, Mathieu, & Zaccaro, 2001).

Much discussion has been devoted to the number of different ways team mental models can be operationalized (e.g., Cooke, Salas, Cannon-Bowers, & Stout, 2000; Mohammed, Klimoski, & Rentsch, 2000). For instance, the similarity or sharedness of team mental models generally describes the degree to which members’ mental models are similar or overlapping. Another characteristic of team mental models is accuracy, which refers to the degree to which members’ mental models adequately represent a given knowledge or skill domain. Researchers (e.g., Stout, Salas, & Kraiger, 1997) have acknowledged that there is sparse research examining measures of team mental model accuracy. Indeed, we were able to locate only two studies (Marks, Zaccaro, & Mathieu, 2000; Webber, Chen, Payne, Marsh, & Zaccaro, 2000) that compared the contributions of similarity and accuracy to team performance. Therefore, our objective in the present study was to examine the relationship between the similarity and accuracy of team mental models and also to compare the unique contribution of each in the prediction of team performance using a longitudinal design. Additionally, consonant with calls for research on effective team composition strategies for the development of team mental models (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000), another objective was to investigate the relationships among team

---

1 In addition to the label mental models, several other labels have been used to describe the construct of knowledge organization such as knowledge structures, schemas, cognitive maps, and conceptual frameworks (Klimoski & Mohammed, 1994).
ability, team mental model similarity and accuracy, and team performance. Finally, we tested the proposition that team mental models would mediate the relationship between team ability and team performance.

Similarity and Accuracy of Team Mental Models

In spite of several calls for the measurement of both the similarity and accuracy of team mental models (e.g., Cannon-Bowers & Salas, 2001; Cooke et al., 2000) surprisingly, the preponderance of the team mental models research (e.g., Converse, Cannon-Bowers, & Salas, 1991; Mathieu et al., 2000; Rentsch & Hall, 1994) has focused only on similarity. We identified three probable explanations for this apparent emphasis on similarity and limited attention to accuracy. First, the emphasis on similarity is consistent with the previous studies’ focus on team process variables because the formation of similar mental models serves as a critical process for achieving effective communication and coordination, which subsequently results in overall improved team performance (Mathieu et al., 2000).

Second, the assessment of team mental model accuracy requires a known “true state of the world” against which a team’s model is compared. However, it would seem that most of the tasks used in the extant literature (e.g., decision making and other tasks for which there is a clearly specified single correct way to perform the task) do not readily lend themselves to generating “true” scores. For instance, instead of using an expert referent model, Marks et al. (2000) had raters judge the accuracy of teams’ concept maps. They argued that there may be multiple correct mental models of the given task, which precludes using a single expert referent model (i.e., “true” score) to represent a single best way to perform the task.

Third, and related to the preceding, measuring accuracy can be a methodological challenge. For instance, Webber et al. (2000) assessed accuracy by comparing team member mental models to an average expert rating that served as the “true” score. However, the measurement and operationalization of mental models by Webber et al. was somewhat different from typical conceptualizations of mental models (e.g., Klimoski & Mohammed, 1994). Specifically, they obtained ratings of the appropriateness of 17 actions or strategies that indicated if they were effective, ineffective, or neutral strategies, but there was no measurement of how these ratings represented relationships among the actions and/or strategies.

Given the issues identified above, we consider the present study to be a constructive replication of previous research because it compared the similarity and accuracy of only taskwork mental models. This is in contrast to Marks et al. (2000) who combined both teamwork (e.g., reporting what other team members are doing) and taskwork (e.g., shooting pillbox, hiding in forest) components in the measurement of team mental models and did not separate them. We also extended the literature by measuring team mental models at two points in time in a longitudinal design, training teams for a longer period of time (i.e., 10 hr over a 2-week interval vs. 1–3 hr), and assessing the effects of team ability composition. We also used a different operationalization of accuracy by comparing trainees’ mental models to an expert referent model that served as the “true” score (see Day, Arthur, & Gettman, 2001, for an example at the individual level).

Contrary to Marks et al. (2000) and Webber et al. (2000), who found a predictive advantage for similarity over accuracy, we posited that there are many tasks for which accuracy would be expected to be a stronger predictor of performance than would similarity. Specifically, accuracy is particularly important in situations or tasks in which there is one best way or a limited set of effective strategies or ways to successfully perform the task—a feature that characterized the task used in the present study. In their analysis of the task used in the present study, Frederiksen and White (1989) identified a single set of internally consistent strategies that distinguish expert performance. The importance of accuracy is based on the rationale that although team members may have shared knowledge, it is possible for this shared knowledge to be inaccurate. As argued by Acton, Johnson, and Goldsmith (1994), when individuals develop expertise, their mental models approximate an expert model and therefore increase in accuracy. Furthermore, as team members develop expertise and their representations converge on the “true” mental model, similarity is expected to increase as accuracy also increases. Thus, team members can have similar but not accurate mental models, although teams with accurate mental models will by definition have similar mental models. So, in some instances, it is plausible that similarity and accuracy could have unique relationships with performance.

A characteristic of past research is that teams have typically been trained for limited amounts of time (i.e., 1–3 hr). Thus, these studies have assessed mental models after very short training sessions. So, as previously noted, it is conceivable that similarity and accuracy will display different patterns of relationships over longer time frames of training (Acton et al., 1994). In summary, on the basis of the preceding conceptual arguments, we tested the following hypotheses.

Hypothesis 1a: The similarity of team mental models will increase over time.

Hypothesis 1b: The accuracy of team mental models will increase over time.

Because there is a clearly defined set of optimal performance strategies for the task used in this study (Frederiksen & White, 1989), we also tested the following hypothesis.

Hypothesis 2: The relationship between team mental model accuracy and team performance will be stronger than the relationship between team mental model similarity and team performance

Team Ability Composition and Team Mental Models

Researchers have stressed the need for more research that describes effective team composition strategies for the development of team mental models (e.g., Mathieu et al., 2000). Kraiger and Wenzel (1997) argued that of all the theoretical antecedents of team mental models, “individual differences are the most proximal variables to shared mental models and thus may have an important impact in their development” (p. 77). Day et al. (2001) showed that ability was related to mental model accuracy for individuals, but no research has examined the relationship between team ability and team mental models.
Given that general mental ability is related to performance through knowledge acquisition (Ree, Carretta, & Teachout, 1995; Schmidt & Hunter, 1992), it is reasonable to posit that team ability should be significantly related to team performance through the development of accurate mental models. That is, teams consisting of members with high mental ability should develop more accurate mental models and consequently have higher team performance than teams with low-ability members. As team members’ mental models converge on the “true” score (i.e., became more accurate), they should also become more similar, such that higher ability teams should also develop similar mental models and have higher team performance than those of lower ability teams. In the present study we manipulated the ability composition of teams by creating teams of uniformly high (HH), mixed (HL), and uniformly low (LL) ability members on the basis of general mental ability. We expected HH teams to have more accurate and similar mental models than HL and LL teams. In addition, we expected that team mental model similarity and accuracy would mediate the relationship between team ability and team performance. Accordingly, we tested the following hypotheses.

Hypothesis 3a: The mental models of HH teams will be more similar than the models of HL teams, which in turn, will be more similar than the models of LL teams.

Hypothesis 3b: The mental models of HH teams will be more accurate than the models of HL teams, which in turn will be more accurate than the models of LL teams.

Hypothesis 4a: The similarity of team mental models will mediate the relationship between team ability composition and team performance.

Hypothesis 4b: The accuracy of team mental models will mediate the relationship between team ability composition and team performance.

Method

Participants

An initial pool of 1,266 male volunteers from Texas A&M University and its community were recruited via advertisements on campus and in local newspapers. Because of hardware constraints, participation was limited to right-handed volunteers. Taking the Raven’s Advanced Progressive Matrices (APM; Raven, Raven, & Court, 1998) as a measure of general mental ability, we used ability scores to create HH, HL, and LL teams. Individuals were retained if they scored 21 or lower or 27 or higher on the APM. These cut scores represented one standard error of measurement above and below the mean APM score on the basis of a standardization sample (N = 496) of college men. This approach ensured that the low- and high-ability participants would indeed be categorically different. At the conclusion of the screening process, 194 individuals were selected and randomly assigned, within ability level, to high-, mixed-, and low-ability teams. Trainees were assigned the same partner throughout training. Twenty-eight of these participants did not complete the study, resulting in an attrition rate of 14%. Chi-square tests indicated no differences in attrition across the three levels of team ability. Thus, the final sample size was 166, which translated into 83 dyadic teams (i.e., 30 HH, 31 HL, and 22 LL dyads). The mean age of the final sample was 19.62 (SD = 2.30).

Participants were paid $75 to participate in a total of 10 days of training. Participants trained for one hr per day on Monday–Friday for 2 consecutive weeks. Participants also had the opportunity to receive a bonus of $50, $30, or $20, which was awarded to each member of the teams with the three highest team scores on the performance task.

Measures

Raven’s APM. The APM is a measure of general mental ability consisting of 36 design problems arranged in an ascending order of difficulty. Because its stimuli are nonverbal and do not require a specific knowledge base to be understood, the APM offers scores that are posited to be uninfluenced by a respondent’s acquired knowledge or reading ability (Saccuzzo & Johnson, 1995). This has led experts to conclude that it is one of the better measures of general mental ability (e.g., Carpenter, Just, & Shell, 1990). We used an administration time of 40 min and obtained a Spearman–Brown odd-even split-half reliability of .84.

Space Fortress. The performance task was the video game Space Fortress (Donchin, 1989; Mané & Donchin, 1989). Space Fortress is “an experimental game which was designed to simulate a complex and dynamic aviation environment” (Gopher, 1993, p. 299). Space Fortress represents important information-processing demands that are present in aviation and other complex tasks (Gopher, Weil, & Bareket, 1994; Hart & Bstatt, 1992). These processing demands include short- and long-term memory load, high workload, dynamic attention allocation, decision making, prioritization, resource management, discrete motor responses, and difficult manual control elements (Gopher, Weil, & Siegel, 1989). Performing Space Fortress involves coordinating mouse and joystick functions to control a spaceship’s flight path and shoot missiles at a fortress. See Arthur et al.’s (1995) article for a more detailed description of Space Fortress.

Pathfinder (Schvaneveldt, 1990; Schvaneveldt, Durso, & Dearholt, 1989). Pathfinder, a computerized structural assessment technique that generates concept similarity maps, was used for the elicitation and analysis of mental models. Pathfinder is a network-scaling procedure (Schvaneveldt, 1990) used to summarize and graphically display relatedness ratings. Pathfinder renders network structures that capture local relationships among concepts. The resulting networks are rich representations that can be quantified and compared (Goldsmith & Davenport, 1990). Two parameters, r and q, determine how network distance is calculated and affect the density of the network. For the present study, the networks were derived with the parameters set to r = infinity and q = the number of concepts (or nodes) minus one.

To generate mental models, we used a set of 14 Space Fortress concepts from Frederiksen and White’s (1989) cognitive task analysis of Space Fortress. Respectively, Table 1 and the Appendix present the concepts and instructions used in the administration of Pathfinder. Trainees made relatedness ratings on all possible pairs of the 14 Space Fortress concepts, which were presented sequentially and randomly to the trainees, resulting in a total of 91 ratings, n(n − 1)/2 = 91, where n = the number of concepts. For each pair of concepts, trainees were asked to indicate the extent to which they were related by using a 9-point Likert scale (1 = not at all related; 9 = highly related).

We used Pathfinder to generate the team mental model indices of similarity and accuracy. Similarity and accuracy were represented by two derivations of the Pathfinder metric of closeness (C), which represents the ratio of common links between two networks divided by the total number of links in both networks. We used C because it is considered to be superior to other Pathfinder metrics such as correlation and number of links (Goldsmith, Johnson, & Acton, 1991; Johnson, Goldsmith, & Teague, 1994) and is also the most commonly used Pathfinder index in the literature. We generated similarity by calculating C between team members’ individual network structures. Team mental model accuracy was defined by the degree of overlap between trainee mental models and an expert referent model. Specifically, we generated accuracy by taking the mean C between each team member’s structure and the expert referent structure. The values
of $C$ for similarity and accuracy range from 0 to 1, with 1 representing
perfect similarity and accuracy.

To obtain the expert referent model, we asked three subject-matter
experts to independently complete Pathfinder. These experts had previously
worked in research labs that used Space Fortress. In addition, they consistently
achieved high scores (i.e., above 4,750) each time they performed Space Fortress. Subsequently, we averaged the three models within the Pathfinder program to yield one referent model, which had an average $C$ of .49. The $Cs$ for each comparison among the three expert models were .46, .58, and .44. Research has indicated that referent models derived from an average of expert judgments yield stronger correlations with performance than those based on a single expert’s judgment (Acton et al., 1994; Day et al., 2001).

### Design and Procedure

Participation involved 10 days of training over a 2-week period. On the
Monday of the first week, trainees began with 20 min of instructions that
explained the rules and optimal strategies of Space Fortress. Trainees then
performed two 3-min baseline Space Fortress games followed by a 5-min
review of the instructions. Trainees then underwent 11 more Space Fortress
training sessions, which took place over the next 13 days. Training sessions
were not scheduled on weekends, and there were 2 training sessions on the
Monday and Friday of the second week. Trainees completed Pathfinder at
the end of Session 1 (i.e., Time 1, which corresponded to the 2nd day of the
10-day training protocol) and Session 3 (i.e., Time 2, which corresponded
to the 4th day of training). Although mental model measurement at the end of
training would have been optimal, we experienced administrative and computer
problems with mental model data collection at Time 3 (i.e., the
9th day of training) that resulted in the loss of data for most teams. Consequently, we have only Time 1 and Time 2 data for all teams. Nevertheless, a 2-day interval between measurements in a 10-day training protocol is still a marked improvement over intervals in other longitudinal designs in which team mental model measurements were taken 20–30 min
apart in a 1–3 hr training protocol (e.g., Marks et al., 2000; Mathieu et al.,
2000).

During a standard training session, the teams performed six practice and
two test games. All games lasted 3 min. For each practice and test game 1
trainee, using his left hand, controlled all functions related to the mouse
(managing mines and missiles), and the other trainee, using his right hand,
controlled all functions related to the joystick and trigger (piloting and
firing the gun). Trainees alternated roles, which called for physically
switching places, at the end of each game. Communication between train-
ees was encouraged. A typical training and testing session lasted approxi-
imately 1 hr, and trainees were scheduled to train at the same 1-hr slot for
their 2 weeks of participation. For each session, performance was opera-
tionalized as the average of the total scores from the two test games.

### Results

**Relationship Between Similarity and Accuracy**

Table 2 presents the descriptive statistics and intercorrelations among all study variables. The correlation between the similarity and accuracy indices was large and statistically significant for both Time 1 ($r = .61, p < .01$, 95% confidence intervals [CIs] = .45–.73) and Time 2 ($r = .67, p < .01$, 95% CI = .53 to .77) measurements. Although team mental model similarity and accuracy increased from Time 1 to Time 2, the increases were not significant: similarity, $t(82) = 1.70, p = .09, d = 0.17$; accuracy, $t(82) = 1.56, p = .12, d = 0.12$. So Hypotheses 1a and 1b were not supported.

**Comparative Criterion-Related Validity of Similarity and Accuracy**

Table 3 reports the means and standard deviations for Space
Fortress team performance for the baseline and all 11 training

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Control ship speed and distance</td>
<td>Maintenance of proper speed and distance from the fortress</td>
</tr>
<tr>
<td>2. Change trajectory of ship</td>
<td>Direction of the ship in frictionless space</td>
</tr>
<tr>
<td>3. Correct press of IFF (mouse)</td>
<td>Accurate identification and resultant response to the appearance of a mine</td>
</tr>
<tr>
<td>4. Incorrect press of IFF (mouse)</td>
<td>Inaccurate identification and resultant response to the appearance of a mine</td>
</tr>
<tr>
<td>5. Select points bonus (mouse)</td>
<td>Choice of this bonus increases points score</td>
</tr>
<tr>
<td>6. Select missiles bonus (mouse)</td>
<td>Choice of this bonus increases missile supply</td>
</tr>
<tr>
<td>7. Friend or foe identifier</td>
<td>Preestablished mine identifiers, shown on the instrument panel</td>
</tr>
<tr>
<td>8. INTRVL (instrument panel)</td>
<td>Interval in milliseconds between mouse button presses indicating when a mine is vulnerable, shown on the instrument panel</td>
</tr>
<tr>
<td>9. Shots counter less than 50</td>
<td>Important information regarding selection of bonus shown on the instrument panel</td>
</tr>
<tr>
<td>10. Shots counter more than 50</td>
<td>Important information regarding selection of bonus shown on the instrument panel</td>
</tr>
<tr>
<td>11. Recognize second $</td>
<td>Important to the acquisition of bonuses</td>
</tr>
<tr>
<td>12. Scoring or losing points</td>
<td>Objective indicator of task performance</td>
</tr>
<tr>
<td>13. Destroy or avoid mines</td>
<td>Handling of friend and foe mines</td>
</tr>
<tr>
<td>14. VLNER (instrument panel)</td>
<td>Number of hits the fortress has suffered that primes the player to apply a “double shot” to destroy the fortress, as shown on the display panel</td>
</tr>
</tbody>
</table>

Note. IFF = identify friend or foe; INTRVL = interval; VLNER = vulnerability.
sessions along with the correlations between the team mental model indices and team performance. These correlations, which are also plotted in Figure 1, indicate that for both Time 1 and Time 2 mental model assessments, the relationships between accuracy and performance were consistently stronger than the relationships between similarity and performance. Indeed, the pattern of results indicates that contrary to the relationships between accuracy and performance, the magnitude of the relationships between similarity and performance decreased with time. Thus, these results provided initial support for Hypothesis 2.

To permit a more succinct and parsimonious presentation of additional tests of Hypothesis 2, we computed the average of the Space Fortress performance scores across Sessions 4–11. We included only Sessions 4–11 in this team performance index because mental models were measured after training Sessions 1 (Time 1) and 3 (Time 2) and we were interested in the predictive criterion-related validity of the mental model indices. In addition, limiting the average team performance to Sessions 4–11 allowed us to hold the time frames of the criterion variable (i.e., team performance) constant to permit interpretable comparisons across the predictors (i.e., team ability, mental model indices). Finally, although team performance substantially improved from Session 1 ($M = 349.66, SD = 1,136.44$) to Session 11 ($M = 3,607.35, SD = 1,733.32$), $t(82) = 20.61, p < .01, d = 2.22$, the decision to average across sessions was deemed appropriate given the magnitude of the intercorrelations among the Space Fortress performance scores ($mean r = .89$).

Consistent with the pattern of results in Figure 1, analyses of the correlations presented in Table 2 indicated that the relations between accuracy and performance (Time 1 $r = .34, p < .01$, 95% confidence interval [CI] $[.21, .47]$) and similarity and performance were stronger at Time 1 than at Time 2 ($r = .32, p < .01$, 95% CI $[.18, .46]$). However, the magnitude of the relationships between similarity and performance decreased with time ($r = .27, p < .01$, 95% CI $[.14, .40]$).

### Table 2
**Descriptive Statistics and Intercorrelations for Study Variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>$M$</th>
<th>$SD$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Team ability</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. Similarity Time 1</td>
<td>.33</td>
<td>.11</td>
<td>.15</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. Accuracy Time 1</td>
<td>.34</td>
<td>.08</td>
<td>.42</td>
<td>.61</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4. Similarity Time 2</td>
<td>.35</td>
<td>.13</td>
<td>.36</td>
<td>.56</td>
<td>.50</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5. Accuracy Time 2</td>
<td>.35</td>
<td>.08</td>
<td>.46</td>
<td>.53</td>
<td>.57</td>
<td>.67</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6. Average team performance</td>
<td>2,999.62</td>
<td>1,592.46</td>
<td>.47</td>
<td>.26</td>
<td>.34</td>
<td>.27</td>
<td>.46</td>
<td>—</td>
</tr>
<tr>
<td>(Sessions 4–11)$^a$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Average team performance</td>
<td>2,139.08</td>
<td>1,374.85</td>
<td>.47</td>
<td>.29</td>
<td>.37</td>
<td>.30</td>
<td>.47</td>
<td>.99</td>
</tr>
<tr>
<td>(All sessions)$^b$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** $N = 83$ dyadic teams. LL = low-ability team; HL = mixed ability team; HH = high-ability team. Team ability was coded as LL = 1, HL = 2, HH = 3. Similarity = team mental model similarity at the given time (Time 1 or Time 2); Accuracy = team mental model accuracy at the given time (Time 1 or Time 2). If $r = .21$ to .28, then $p < .05$; if $r > .29$, then $p < .01$.

$^a$ Average team performance was the average of Sessions 4–11. $^b$ Average team performance was the average of the baseline and Sessions 1–11; this information is presented for the sake of completeness.

### Table 3
**Descriptive Statistics for Space Fortress Team Performance and Correlations Between Team Mental Model Indices and Team Performance**

<table>
<thead>
<tr>
<th>SF session</th>
<th>$M$</th>
<th>$SD$</th>
<th>Similarity</th>
<th>Accuracy</th>
<th>Similarity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time 1</td>
<td>Time 1</td>
<td>Time 2</td>
<td>Time 2</td>
</tr>
<tr>
<td>Baseline</td>
<td>−1,854.29</td>
<td>713.49</td>
<td>.26</td>
<td>.29</td>
<td>.27</td>
<td>.32</td>
</tr>
<tr>
<td>1</td>
<td>349.66</td>
<td>1,136.44</td>
<td>.29</td>
<td>.35</td>
<td>.28</td>
<td>.36</td>
</tr>
<tr>
<td>2</td>
<td>1,228.93</td>
<td>1,352.44</td>
<td>.32</td>
<td>.38</td>
<td>.35</td>
<td>.40</td>
</tr>
<tr>
<td>3</td>
<td>1,947.71</td>
<td>1,528.90</td>
<td>.30</td>
<td>.38</td>
<td>.27</td>
<td>.45</td>
</tr>
<tr>
<td>4</td>
<td>2,100.14</td>
<td>1,487.38</td>
<td>.30</td>
<td>.34</td>
<td>.33</td>
<td>.45</td>
</tr>
<tr>
<td>5</td>
<td>2,452.04</td>
<td>1,540.87</td>
<td>.30</td>
<td>.36</td>
<td>.24</td>
<td>.43</td>
</tr>
<tr>
<td>6</td>
<td>2,717.27</td>
<td>1,692.89</td>
<td>.23</td>
<td>.29</td>
<td>.26</td>
<td>.44</td>
</tr>
<tr>
<td>7</td>
<td>3,027.35</td>
<td>1,695.94</td>
<td>.23</td>
<td>.33</td>
<td>.27</td>
<td>.44</td>
</tr>
<tr>
<td>8</td>
<td>3,148.32</td>
<td>1,928.13</td>
<td>.25</td>
<td>.31</td>
<td>.22</td>
<td>.44</td>
</tr>
<tr>
<td>9</td>
<td>3,432.21</td>
<td>1,657.50</td>
<td>.24</td>
<td>.31</td>
<td>.26</td>
<td>.42</td>
</tr>
<tr>
<td>10</td>
<td>3,512.24</td>
<td>1,636.10</td>
<td>.23</td>
<td>.30</td>
<td>.26</td>
<td>.45</td>
</tr>
<tr>
<td>11</td>
<td>3,607.35</td>
<td>1,733.32</td>
<td>.25</td>
<td>.37</td>
<td>.25</td>
<td>.44</td>
</tr>
<tr>
<td>Average team performance</td>
<td>2,999.62</td>
<td>1,592.46</td>
<td>.26</td>
<td>.34</td>
<td>.27</td>
<td>.46</td>
</tr>
<tr>
<td>(Sessions 4–11)$^a$</td>
<td>2,139.08</td>
<td>1,374.85</td>
<td>.28</td>
<td>.37</td>
<td>.30</td>
<td>.47</td>
</tr>
</tbody>
</table>

**Note.** All correlations were significant at $p < .05$. SF = Space Fortress.

$^a$ Average team performance was the average of Sessions 4–11. $^b$ Average team performance was the average of the baseline and Sessions 1–11; this information is presented for the sake of completeness.
CI = .13 to .52; Time 2 $r = .46$, $p < .01$, 95% CI = .27 to .61) were stronger than those between similarity and performance (Time 1 $r = .26$, $p < .05$, 95% CI = .05 to .45; Time 2 $r = .27$, $p < .01$, 95% CI = .06 to .46). A test for differences between dependent $r$s revealed that although the difference between the similarity and accuracy correlations with performance was not significant at Time 1, $t(80) = 0.86$, ns, it was significant at Time 2, $t(80) = 2.36$, $p < .05$, lending additional support for Hypothesis 2.

To investigate the incremental variance explained in team performance by similarity and accuracy, we computed four hierarchical regression equations (see Table 4). In all models, we entered team ability composition in the first step of the hierarchical regression. In the first model, we entered similarity in the second step, followed by accuracy in the third step to examine the incremental validity of accuracy beyond team ability and similarity measured at Time 1. In the second model, we reversed the entry order of similarity and accuracy to examine the unique contribution of similarity beyond team ability and accuracy measured at Time 1. In the first model, we reversed the entry order of similarity and accuracy to examine the unique contribution of similarity beyond team ability and similarity measured at Time 1. Models 3 and 4 were replications of the first two models using similarity and accuracy measured at Time 2. The results revealed that similarity composition accounted for 22% of the variance in team performance. The results of Models 1 and 2 indicated that neither accuracy (Model 1 $\Delta R^2 = .00$, ns) nor similarity (Model 2 $\Delta R^2 = .02$, ns) explained unique variance in team performance beyond team ability and the other mental model index. However, for Time 2 data, accuracy explained unique variance in team performance beyond team ability and similarity (Model 3 $\Delta R^2 = .07$, $p < .01$), but similarity did not explain unique variance in team performance (Model 4 $\Delta R^2 = .01$, ns).

![Figure 1](image.png)

**Figure 1.** Correlations between team similarity and accuracy measured at Time 1 and Time 2, and team performance measured at the baseline and Sessions 1–11; Sim T1 = Similarity measured at Time 1. Acc T1 = Accuracy measured at Time 1. Sim T2 = Similarity measured at Time 2; Acc T2 = Accuracy measured at Time 2. B = Baseline performance session.

Table 4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>$R^2$</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team ability</td>
<td>.47**</td>
<td>.44**</td>
<td>.41**</td>
<td>.22**</td>
<td></td>
</tr>
<tr>
<td>Mental model similarity</td>
<td>.20*</td>
<td>.16</td>
<td>.26**</td>
<td>.04*</td>
<td>.00</td>
</tr>
<tr>
<td>Mental model accuracy</td>
<td></td>
<td>.07</td>
<td>.26**</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team ability</td>
<td>.47**</td>
<td>.39**</td>
<td>.41**</td>
<td>.22**</td>
<td></td>
</tr>
<tr>
<td>Mental model accuracy</td>
<td>.17</td>
<td>.07</td>
<td>.24**</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>Mental model similarity</td>
<td>.16</td>
<td>.26**</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team ability</td>
<td>.47**</td>
<td>.42**</td>
<td>.33**</td>
<td>.22**</td>
<td>.01</td>
</tr>
<tr>
<td>Mental model similarity</td>
<td>.12</td>
<td>.10</td>
<td>.23**</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>Mental model accuracy</td>
<td>.37**</td>
<td>.30**</td>
<td>.07**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team ability</td>
<td>.47**</td>
<td>.32**</td>
<td>.33**</td>
<td>.22**</td>
<td>.01</td>
</tr>
<tr>
<td>Mental model accuracy</td>
<td>.31**</td>
<td>.37**</td>
<td>.29**</td>
<td>.07**</td>
<td></td>
</tr>
<tr>
<td>Mental model similarity</td>
<td>−.10</td>
<td>.30**</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $N = 83$ dyadic teams. LL = low-ability team; HL = mixed-ability team; HH = high ability team. Team ability was coded as LL = 1, HL = 2, HH = 3. The dependent variable was average team performance operationalized as the average of Sessions 4–11.

* $p < .05$. ** $p < .01$. 

*732 RESEARCH REPORTS*
Team Ability, Mental Models, and Team Performance

Hypothesis 3a predicted that the mental models of HH teams would be more similar than mental models of HL teams, which in turn would be more similar than the models of LL teams. Hypothesis 3b made the same predictions for accuracy. The results in Table 5 show that the similarity and accuracy of team mental models differed across the three ability compositions in the predicted direction. A 3 (team ability) × 2 (mental models administration) analysis of variance showed a significant main effect for ability, F(2, 80) = 3.98, p < .05, η² = .09, no significant main effect for mental models administration, F(1, 80) = 1.79, ns, η² = .00, and a significant interaction, F(2, 80) = 3.22, p < .05, η² = .02. Paired comparisons showed that there were no differences between the ability compositions for similarity measured at Time 1. However, for Time 2 there were significant differences for the HH–LL comparisons, t(50) = 3.70, p < .01, d = 0.96 and the HL–LL comparisons, t(51) = 2.34, p < .05, d = 0.58, although the HH–HL difference was not statistically significant, t(59) = 1.28, ns, d = 0.35. These results indicated that mental model similarity did not differ by team ability level when measured at Time 1, but after two more sessions of training, the HH and HL teams had developed more similar mental models than had the LL teams.

Regarding Hypothesis 3b, a 3 × 2 analysis of variance for accuracy showed a significant main effect for ability, F(2, 80) = 13.09, p < .01, η² = .32, but the main effect for mental models administration, F(1, 80) = 2.37, ns, η² = .00, and the interaction, F(2, 80) = 0.44, ns, η² = .00, were not significant. Paired comparisons showed that for Time 1 there were significant differences in mental model accuracy among all three ability compositions: HH–LL, t(50) = 3.94, p < .01, d = 1.08, HL–LL, t(51) = 2.54, p < .01, d = 0.73, HH–HL, t(59) = 1.91, p < .05, d = 0.46. Comparisons for the accuracy index measured at Time 2 also revealed statistically significant differences among all three team ability compositions: HH–HL, t(59) = 2.62, p < .01, d = 0.67; HL–LL, t(51) = 2.31, p < .05, d = 0.54; HH–LL, t(50) = 4.36, p < .01, d = 1.13. In general, the pattern of results demonstrated support for Hypotheses 3a and 3b in that the similarity and accuracy of team mental models differed among all three team ability compositions. The only exceptions to this pattern of results were the nonsignificant effects for similarity measured at Time 1 and the difference between the HH and HL teams for similarity measured at Time 2.

Hypotheses 4a and 4b stated that mental models (similarity and accuracy, respectively) would mediate the relationship between team ability and team performance and were tested in accordance with standards outlined by Baron and Kenny (1986). We conducted the tests of mediation for similarity and accuracy separately, using mental models data collected at Time 2 because we considered these data to be more robust as they encompassed more training. The results of the similarity mediation test, which are presented as Model 3 in Table 4, indicate a small decrease in the effect of team ability after controlling for mental model similarity (β = .47 vs. β = .42). However, contrary to Hypothesis 4a, Sobel’s (1982) test showed that the indirect effect of team ability and team performance through mental model similarity was not significantly different from 0, t(81) = 1.09, ns; indirect effect = 88.83, 95% CI = –70.67 to 248.34. The results of the accuracy mediation test, which are presented as Model 4 in Table 4, indicate a decrease in the effect of team ability when controlling for mental model accuracy (β = .47 vs. β = .32). An examination of the 95% CI for the indirect effect of team ability and team performance through mental model accuracy indicates that the effect was significantly different from 0, indirect effect = 284.40, 95% CI = 54.75 to 514.04; t(81)= 2.43, p < .05. Therefore, in support of Hypothesis 4b, mental model accuracy partially mediated the relationship between team ability and team performance.

Table 5
Descriptive Statistics for Mental Models and Team Performance of the Three Team Ability Compositions

<table>
<thead>
<tr>
<th>Variable</th>
<th>HH</th>
<th></th>
<th>LL</th>
<th></th>
<th>F(2, 80)</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team abilitya</td>
<td>29.98</td>
<td>1.85</td>
<td>23.95</td>
<td>1.99</td>
<td>18.45</td>
<td>1.88</td>
</tr>
<tr>
<td>Similarity Time 1</td>
<td>.35</td>
<td>.12</td>
<td>.33</td>
<td>.11</td>
<td>.30</td>
<td>.10</td>
</tr>
<tr>
<td>Accuracy Time 1</td>
<td>.37</td>
<td>.07</td>
<td>.34</td>
<td>.06</td>
<td>.29</td>
<td>.08</td>
</tr>
<tr>
<td>Similarity Time 2</td>
<td>.39</td>
<td>.11</td>
<td>.35</td>
<td>.12</td>
<td>.28</td>
<td>.12</td>
</tr>
<tr>
<td>Accuracy Time 2</td>
<td>.39</td>
<td>.08</td>
<td>.34</td>
<td>.07</td>
<td>.30</td>
<td>.08</td>
</tr>
<tr>
<td>Average team performance</td>
<td>3,924.82</td>
<td>1,303.20</td>
<td>2,761.11</td>
<td>1,591.72</td>
<td>2,074.05</td>
<td>1,316.54</td>
</tr>
<tr>
<td>(Sessions 4–11)b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average team performance</td>
<td>2,968.46</td>
<td>1,158.79</td>
<td>1,895.65</td>
<td>1,372.67</td>
<td>1,351.10</td>
<td>1,055.51</td>
</tr>
<tr>
<td>(All sessions)c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. HH = two high-ability members (n = 30 dyadic teams); HL = one high- and one low-ability member (n = 31 dyadic teams); LL = two low-ability members (n = 22 dyadic teams). The F values and eta squared were generated from five one-way analyses of variance using team ability composition as the independent variable and each index and average team performance as the dependent variables.

*a Team ability was the average Advanced Progressive Matrices scores of members within each ability composition. b Average team performance was the average of Sessions 4–11. c Average team performance was the average of the baseline and Sessions 1–11; this information is presented for the sake of completeness.

**p < .01.
Discussion

Our findings support a growing body of research that indicates team mental models play an important role in the development of complex skills and subsequent team performance (Marks et al., 2000; Mathieu et al., 2000). Previous research has primarily focused on the measurement of team mental model similarity at the exclusion of accuracy (e.g., Levesque, Wilson, & Wholey, 2001; Mathieu et al., 2000; Peterson, Mitchell, Thompson, & Burr, 2000; Rentsch & Klimoski, 2001). Therefore, one of our objectives in the present study was to compare the similarity and accuracy of team mental models in a longitudinal research design.

Although our results did not show a significant increase in similarity and accuracy over time—most likely because of the relatively short time interval between the two mental model administrations—the similarity and accuracy of team mental models were strongly related. We also showed that although the similarity and accuracy of mental model indices taken early in team training (i.e., at Time 1) were equally predictive of team performance, after 4 days of training (i.e., at Time 2), the accuracy of team mental models was a stronger predictor of subsequent team performance. Prior research on team mental models tends to favor similarity as the stronger predictor of team performance, but these data have been typically collected in designs with 1–3 hr training and performance sessions (e.g., Marks et al., 2000; Mathieu et al., 2000).

Thus, the differences in our findings and those previously reported for similarity might be due to our use of a longer team-training time frame. Our longitudinal data suggest that we would have obtained results similar to those reported in the extant literature if we had terminated our data collection at the end of the first mental model assessment (i.e., after 2 hr of training and performance). However, our results (e.g., Figure 1) clearly show that as teams acquired more skill and converged on the “true” mental model with increased training (Acton et al., 1994), the comparative validity of similarity and accuracy changed, with accuracy becoming a stronger predictor than similarity.

In contrast to previous research (e.g., Marks et al., 2000; Webber et al., 2000), the present study used a task in which there was a limited number of effective strategies and focused on only taskwork mental models. These features allowed us to obtain an expert referent mental model that served as the “true” score and subsequently operationalize team mental model accuracy as the degree of overlap between trainee mental models and an expert referent model. The differences in tasks and our use of longer training and performance time frames may be plausible explanations for the differences in our results and those of Marks et al. (2000), who used a decision-making task and 3 hr of training and performance and found stronger effects for mental model similarity.

Another contribution of the present study is its investigation of the relationships among team ability, team mental model similarity and accuracy, and team performance. Specifically, our results indicate that the similarity and accuracy of team mental models are related to team general mental ability. However, team ability is more strongly related to the accuracy than to the similarity of team mental models. Thus, we demonstrated that team ability is an important predictor of the accuracy and, to a lesser extent, the similarity of team mental models.

However, because we focused exclusively on taskwork mental models, our data do not speak to the role of team ability in the development of other forms of knowledge organization, such as teamwork models. It is conceivable that the comparative validity of similarity and accuracy may be a function of whether the focus is on taskwork versus teamwork. Nevertheless, we demonstrated that when the focus is on taskwork mental models, accuracy of mental models is a better predictor of team performance than is their similarity. Now that we have established this boundary condition for taskwork mental models, future research could focus on measuring taskwork and teamwork mental models separately in a single study.

We also demonstrated that the accuracy of team mental models partially mediates the relationship between team ability and team performance. General mental ability is related to performance through knowledge acquisition (Schmidt & Hunter, 1992). Given that mental models are representations of knowledge in a given domain, it is not surprising that HH teams developed more accurate mental models and subsequently higher team performance.

Although a strength of the present study is that teams participated in a much longer training protocol (2 weeks) than in most team training laboratory studies (typically 1–3 hr), the external validity of our findings may still be somewhat restricted because of the limited life span of our teams. In spite of this, the results of the present study are most likely to generalize to teams that perform tasks for which there is a demonstrable best or limited set of effective strategies. In contrast, for tasks with multiple correct ways or effective strategies, there are likely to be multiple, accurate team mental models. Consistent with the concept of equifinality, it would be difficult, if not impossible, to determine the definitive accurate mental model. Consequently, under these conditions, similarity (and teamwork mental models) may be more important than accuracy.

Also, although we used dyadic teams and larger teams have more complex dynamics than dyads, it is not unreasonable to posit that our findings may generalize to larger teams responsible for tasks similar to the one used in the present study; that is, tasks for which there is a limited number of optimal strategies. However, further research is needed to test this proposition. Finally, our results have additional implications for research and the training and development of teams in the field. First, they would suggest that wherever possible, one could train for accuracy with the expectation that similarity would follow. Second, where it is possible to generate expert referent mental models, one could investigate their efficacy as interventions to facilitate the development of accurate and shared team mental models. Third, the mental models of trainees could be assessed during training, and trainers could use the expert model as a means of providing corrective feedback. Fourth, our results suggest that when feasible one can influence team mental models, and subsequently team performance by manipulating team composition in terms of team general mental ability. Fifth, whereas the measurement of mental models in the field may be an administrative challenge, there is no reason why the processes would be any different from measurements used in the laboratory as long as the task concepts can be explicitly defined (e.g., Smith-Jentsch, Campbell, Milanovich, & Reynolds, 2001).
Conclusion

We presented evidence that for a task with a defined set of optimal strategies, team mental model accuracy is a stronger predictor of team performance than team mental model similarity. However, unlike previous research that tends to favor similarity, this pattern of results did not emerge until later in training. In response to calls for the exploration of the determinants of team mental models, the present study also provides evidence that team members’ ability is related to the development of similar and accurate mental models and that the accuracy of mental models partially mediates the relationship between team ability and team performance.

References


**Appendix**

Instructions for Making Relatedness Ratings in Pathfinder

Your task on the computer is to make judgments about the “relatedness” of pairs of terms that have to do with playing the Space Fortress game. There are several ways one might think about the terms being judged. For instance, two terms might be related because they share common features or because they frequently occur together. For this task think about the terms as they relate to playing Space Fortress.

YOU SHOULD BASE YOUR RELATEDNESS JUDGMENTS ON HOW THE TERMS WORK TOGETHER TO HELP YOU PLAY THE GAME WELL.

The major goal of Space Fortress is to maximize your game points. This is accomplished by: (1) Destroying the Fortress as many times as you can, (2) Hitting as many mines as possible, and (3) Protecting your own ship from being hit or damaged. Thus, when making your relatedness judgments you should think about the Space Fortress in relation to these 3 sub-goals.

Each pair of terms will be presented on the screen along with a “relatedness” scale. You can think of the points along the scale as representing degrees of relatedness ranging from “1”, not at all related to “9”, highly related.

You are to indicate your judgement of relatedness for each pair of terms by pressing the number key that represents your rating. Upon responding, a bar marker will move directly above the number you pressed. If you wish to change your rating, simply press another number. When you are satisfied with the rating you have given, press the <SPACE BAR> to enter your response. Following this, the next pair of terms will be displayed.

A complete list of terms will be presented prior to beginning the task. This will give you a general idea of the scope of the Space Fortress terms you will be rating.

WHEN MAKING YOUR RATINGS, REMEMBER BACK TO PLAYING THE SPACE FORTRESS GAME AND THINK ABOUT HOW THE TERMS ARE RELATED IN ACCOMPLISHING THE GOAL OF MAXIMIZING YOUR SCORE.

Received April 30, 2004
Revision received February 24, 2005
Accepted March 10, 2005