Observational learning is based on a critical assumption that trainees can and do recognize critical modeled behaviors. This assumption has been virtually untested in applied settings. We studied the effects of work experience and instructions on the ability of 59 observers to recognize target behaviors in an observational learning paradigm similar to existing ones. Additionally, we investigated the effects of two key factors that were hypothesized to affect the recognition process in observational learning. The results indicated that only observers who had a minimum of work experience (i.e., intermediate and experienced observers in the study) were able to consistently recognize targeted behaviors. Additionally, recognition was influenced by the level of detail of instructions given to the participants. Finally, characteristics of the modeled behaviors greatly affected recognition: Overall, examples of negative behaviors were better recognized than were positive examples. Behaviors whose consequence was shown were also better recognized than those that were neither reinforced nor punished in the video. The results are discussed in terms of their implications for the design of observational learning as a training strategy in complex and applied social learning situations. The applications of this work include the design of training, and the training of evaluators and observers.
faced by trainees trying to learn (Borman & Hallam, 1991).

That so little attention has been paid to the factors that might influence trainees’ attention processes regarding observational learning stimuli may be partly attributable to the success of observational learning in empirical studies. That is, because observational learning works, trainees must be able to recognize critical elements. It might also be the case that previous studies have used observational learning in environments where the salience of training cues is very high. Thus the issue of attention would not be critical. However, observational learning is now being applied in relatively complex social situations in which training cues (and consequences) exist in a complicated stimulus array. For example, observational learning has been used to train managerial and communication skills to managers (e.g., Nunns & Bluen, 1992), pilots (Smith-Jentsch, Jentsch, Payne, & Salas, 1996), and others (Smith-Jentsch, Salas, & Baker, 1996). In these areas the salience of targeted behaviors might be a critical factor in training effectiveness (Mann & Decker, 1984; Thelen & Rennie, 1972).

The limited literature on observational learning indicates that three groups of factors in particular can be hypothesized to affect the recognition of targeted behaviors in modeling displays. The first involves individual differences in task-relevant knowledge and experiences. Because of their greater knowledge bases, experts likely are better able to recognize examples of appropriate and inappropriate behaviors than are novices. Displays that are suitable for experts may therefore be unsuitable for task novices.

The second group of factors includes the instructions given to observers. Providing observers with a mental model or schema before observation, for example, has been found to positively affect transfer performance, because such schemas allow observers to classify and organize their expectations (e.g., Trimble, Nathan, & Decker, 1991). The role that recognition of the targeted behaviors played in achieving these effects, however, was not studied. Yet it seems reasonable to hypothesize that attention processes in observational learning can be augmented by detailed instructions about the targeted behaviors.

According to social cognitive theory, observational learning requires engagement in cognitive processes similar to those in actual practice or experience. This notion was recently supported by research on the processes involved in observational learning of psychomotor tasks (Blandin, Lhuisset, & Proteau, 1999; Blandin & Proteau, 2000). Thus a case can be made that novices should benefit more from specific instructions than should experts because they especially require a schema to classify and organize their expectations. This would suggest that expertise and preobservation instructions should interact to facilitate recognition of targeted displays.

Third, characteristics of the targeted behaviors may interact with the aforementioned factors. Two main characteristics of modeled behaviors are the use of positive versus negative examples and whether or not a consequence is shown for a modeled behavior. With respect to the former variable, Baldwin (1992) demonstrated that trainees who observed only positive models were better able to reproduce the targeted behaviors than were those who had watched positive and negative examples. The latter, in turn, were better able to generalize the targeted behaviors in a transfer test. Again, however, it is not entirely clear from this research whether recognition was significantly affected by such changes or whether the effects of varying the display characteristics occurred at a later point in learning.

Finally, with respect to the question of whether consequences should be shown for a modeled behavior, social learning and script-processing theories predict that showing the consequences of a behavior should not only increase its salience but also foster the establishment of critical antecedent-behavior-consequence links that assist observers in translating what is watched into rules for their own behavior (Gioia & Manz, 1985).

The purpose of the present study was, therefore, to investigate the effects of four factors on recognition of target behaviors in a modeling display: (a) observer experience, (b) preobservation instructions, (c) type of behavior shown (positive/negative), and (d) whether or not a consequence for a behavior was shown. Pilots from three experience levels
were randomly assigned to three experimental conditions, which varied according to the specificity of preobservation instructions provided to the participants. The pilots then watched a videotaped flight simulation that contained a number of positive and negative examples of the targeted behaviors, some for which consequences were shown and others for which no consequences were shown.

It was expected that recognition of targeted behaviors would be greatest in experienced aviators than in less experienced aviators. Further, we expected that the specificity of the instructions affected how many behaviors would be recognized and how well they would be described by observers. Recognition was also hypothesized to be positively influenced by factors thought to increase cue salience. Specifically, we anticipated that high-salience behaviors (i.e., positive examples and those whose consequences were shown) would be recognized more readily than lower-salience cues (i.e., negative examples and those whose consequences were not shown).

METHOD

Participants

Participants were 60 pilots from an airline-affiliated flight academy who volunteered for the study. In addition to 16 flight instructors (mean total time [TT] = 753 hr), 11 commercial pilots (mean TT = 256 hr), 23 private pilots (mean TT = 186 hr), and 10 student pilots (mean TT = 33 hr) participated. Of the 60 pilots, 32 held an instrument flight rules rating and 28 held a visual flight rules rating.

Participants were divided into three experience groups, according to a combination of pilot certificate and the minimum flight time required for its unrestricted use. The 15 student or private pilots with fewer than 120 hours of total flight time (i.e., the minimum time for an instrument rating) were assigned to the low-experience group (LE). The medium-experience group (ME) was made up of the 25 pilots with either a private pilot license and 120 hours or more of total flight experience as a pilot, or a commercial pilot license and fewer than 250 hours. (Inspection of the data revealed that one pilot in the ME group did not complete all data collection instruments and had to be removed from the analyses. This left a total of 59 pilots in the data set for all analyses.) Finally, the 22 pilots with unrestricted commercial pilot, flight instructor, or airline transport pilot licenses and more than 250 hours of total flight time formed the high-experience group (HE).

The pilots from each experience level were randomly assigned to one of the three preobservation instructions conditions, as equally as possible given that there were not even multiples of 3 in all experience groups. Thus, 5 (± 1) pilots from the low-experience group were in each of the three preobservation instructions conditions, as well as 7 (+ 1) from the medium-experience group, and 7 (+ 1) from the high-experience group, respectively.

The three preobservation instructions conditions varied according to the information given to the pilots before the observation exercise. In the first preobservation instructions condition (OBS CREW), the pilots were told only that they should observe and record positive and negative crew behaviors – that is, they were neither told that they should only observe the copilot, nor were they given specific instructions as to which behaviors they should look for. In the second preobservation instructions condition (OBS COPILOT), pilots were made specifically aware that they should observe and record positive and negative behaviors exhibited by the copilot. Finally, pilots in the third preobservation instructions condition (OBS SPEC COPILOT) were told that they should observe copilot backup behaviors and record positive and negative examples of behaviors in six behavioral categories: (a) planning ahead, (b) providing information in advance, (c) distributing tasks to avoid overloading, (d) “buying the flight some time” for decisions, (e) pointing out problems or discrepancies, and (f) recording information in permanent form. Table 1 shows the instructions associated with the three preobservation instructions conditions.

Design

The study used a 3 (Experience) × 3 (Pre-observations-Instruction) × 2 (Positive/Negative) × 2 (Consequence-Shown/Not-Shown) mixed-
model design. Between-subjects factors were experience with three levels (low, medium, high) and preobservation instructions with three levels (observe crew, observe copilot, observe a set of specific set of behaviors for the copilot). Within-subjects factors were two salience cues: quality of the displayed behavior (positive/negative example) and whether a consequence was shown (yes/no).

**Procedure**

After participants were informed about the purpose of the study and assured confidentiality of their data, they completed an informed consent form. They then filled out a number of biodata measures. After a short break, the pilots were given instructions for the observation in accordance to their experimental group.

The pilots watched a 45-minute videotape of a flight scenario that contained everything from the first meeting of the pilots and their briefing through engine shutdown at the destination. The scenario was scripted as a small jet operation from San Juan, Puerto Rico, to St. Croix in the U.S. Virgin Islands. This route was selected because it was reasonable to assume that most participants were unfamiliar with the route and the local conditions; a fact later supported by postvideotape questioning. Two Reserve naval aviators followed a script and acted as the crew for the fictional Bluebird Airlines.

A number of technical, procedural, and management behaviors were scripted in accordance with the experimental design. Seventeen scripted situations modeled the six classes of copilot behaviors that formed the focus of the study. The conflicts varied with respect to the quality of the displayed behaviors (i.e., whether a positive or a negative example was shown) and to whether a consequence for the behavior was shown (yes/no). For about half the 17 targeted behaviors, a consequence of the copilot’s behavior was shown (i.e., the captain or air traffic control either punished or rewarded the behavior). For the other behaviors, however, such an “external” consequence was not shown. For example, in one situation, the crew was flying at low altitude toward a mountain, but nothing happened because the flight was diverted for a reason other than terrain clearance before hitting the mountain.

The pilots recorded their observations on worksheets to indicate whether the observed behavior was positive or negative; they were also asked to provide a brief description of the observed behavior. After the videotape ended, these sheets were collected, and all pilots were asked to write on a second worksheet as many specific examples of positive backup behaviors by a copilot as they could. This was a simple generalization measure. After a brief break, participants were given written feedback on a form containing descriptions of all 17 targeted

---

**TABLE 1: Instructions for the Three Sets of Preobservation Instructions**

<table>
<thead>
<tr>
<th>Preobservation Instructions</th>
<th>OBS CREW</th>
<th>OBS COPilot</th>
<th>OBS SPEC COPilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watch and record positive and negative examples of the crew’s behaviors.</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Watch and record specifically positive and negative examples of the copilot’s behaviors.</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Watch and record specifically the behaviors of the copilot that fall into any of the following six behavioral categories:</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Planning ahead</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Providing information in advance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) Distributing tasks to avoid overloading</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d) “Buying the flight some time” for decisions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e) Pointing out problems of discrepancies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f) Recording information in permanent form</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
behaviors and an explanation of the associated category of backup behaviors. We asked participants to indicate whether they had noticed the behavior ("yes"/"no") and how distinctive/recognizable they considered it (on a 7-point Likert scale).

After the feedback sheets had been collected, the pilots were again asked to generate as many examples of positive copilot backup behaviors as possible. The experimental sessions ended with a test of declarative knowledge in aviation and a reactions questionnaire.

**Measures**

The following measures were employed in the study. Descriptive statistics (means, standard deviations, and alphas, where appropriate) and intercorrelations are provided in Table 2.

**Biodata.** Paper-and-pencil measures were used to collect biodata from the pilots. Data were collected regarding age, highest certificates, total flight time, future plans, and reasons for participating in the study.

**Measure of declarative knowledge.** The pilots completed a measure of aeronautical declarative knowledge consisting of 20 items from the Federal Aviation Administration’s (FAA) test bank for the instrument rating test. This took the form of three-option multiple-choice tests.

**Recognition measures.** The pilots recorded their observations on a recognition sheet that was rated by a trained pilot and flight instructor who served as a condition-blind rater. The pilot used two rating criteria to categorize the answers. Using the first criterion, the rater merely decided whether the targeted behavior was mentioned at all ("1") or not ("0"). Because 1 of the 17 targeted behaviors was recognized by only 1 of the 60 participants, only the 16 remaining behaviors were used (\(\alpha = .54\)). The rater also used a six-point anchored rating scale to indicate the quality of the description given by the pilots. Scores of 0 corresponded to no recognition at all; a 5 indicated that the example was recognized, described, and related to the underlying category of backup behaviors (\(\alpha = .81\)). All answer sheets were then controlled-rated by a second condition-blind rater using the same criteria. Agreement among the two raters as indicated by Pearson \(r\) was \(r = .81\) for the dichotomous ratings and \(r = .82\) for the scale ratings.

### Table 2: Means, Standard Deviations, Internal Consistencies (Cronbach’s Alpha), and Intercorrelations for the Measures Used in the Study

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Flight time (hr)</td>
<td>325.00</td>
<td>339.60</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Declarative knowledge (% on FAA written test)</td>
<td>70.00</td>
<td>18.75</td>
<td>.28*</td>
<td>(.77)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Number of conflicts recognized (0–100%)</td>
<td>55.30</td>
<td>14.19</td>
<td>.13</td>
<td>.33*</td>
<td>(.54)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Quality of conflicts (0–5) descriptions</td>
<td>1.70</td>
<td>0.78</td>
<td>.15</td>
<td>.25</td>
<td>.89**</td>
<td>(.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Distinctiveness ratings (0–7)</td>
<td>5.68</td>
<td>0.71</td>
<td>−.22</td>
<td>.06</td>
<td>.15</td>
<td>.25</td>
<td>(.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Number of items generated before feedback</td>
<td>11.33</td>
<td>3.90</td>
<td>.14</td>
<td>.21</td>
<td>.51**</td>
<td>.49**</td>
<td>.22</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>7. Number of items generated after feedback</td>
<td>9.83</td>
<td>4.25</td>
<td>.04</td>
<td>.01</td>
<td>.17</td>
<td>.06</td>
<td>.13</td>
<td>.37**</td>
<td>–</td>
</tr>
</tbody>
</table>

\(N = 59, \ *p < .05 \) (two-tailed), \ **p < .01 \) (two-tailed), numbers in parentheses indicate coefficient alpha, where appropriate.
Distinctiveness measure. In the feedback phase, participants were informed about the targeted behaviors and then asked to rate each one, on a 7-point Likert scale (1 = not at all, 7 = very), on how distinctive they found it. Cronbach’s alpha for the ratings was .80.

Generalization measures. One simple generalization measure – the number of positive examples of copilot backup behaviors the pilots could generate – was taken before and after feedback.

RESULTS

An alpha level of .05 was used for all analyses.

Manipulation Checks

Using planned comparisons, all manipulation checks were significant. Pilots in the lowest experience group had significantly fewer flight hours ($M_{LE} = 52$ h) than did those in the medium experience group ($M_{ME} = 216$ h), $t(56) = 1.97$, $p_{1-tail} = .027$. Medium-experience pilots, in turn, had significantly fewer flight hours than did those in the highest experience group ($M_{HE} = 625$ h), $t(56) = 5.47$, $p_{1-tail} < .001$. Further, as expected, pilots in both the ME and HE groups had significantly higher average scores on the FAA instrument written test, $M_{ME} = 78\%$, $t(56) = 6.73$, $p_{1-tail} < .001$; and $M_{HE} = 77\%$, $t(56) = 6.33$, $p_{1-tail} < .001$, respectively, than did those in the LE group, $M_{LE} = 48\%$.

The effectiveness of the preobservation instructions manipulation was also significant: As expected, pilots who were told to only observe the copilot reported proportionally more copilot behaviors, $M_{OBS COPILOT} = 84\%$ and $M_{OBS SPEC COPILOT} = 84\%$, respectively, than did those who were not, $M_{OBS CREW} = 27\%$, $t(56) = -15.12$, $p_{1-tail} < .001$. Further, as expected, participants who were given the examples of backup behaviors recorded a significantly higher proportion of copilot behaviors that were related to the list, $M_{OBS SPEC COPILOT} = 41\%$, than did the pilots who were not given the list, $M_{OBS CREW} = 35\%$ and $M_{OBS COPILOT} = 30\%$, respectively, $t(56) = 2.17$, $p_{1-tail} = .017$.

Analysis Strategy

Two independent 3 (Preobservation Instructions) × 3 (Pilot Experience) × 2 (Consequence Shown/Not Shown) × 2 (Positive/Negative Example) mixed-model ANOVAs were calculated on the percentage of behaviors recognized by the pilots and the quality scores of their descriptions. Additionally, a similar 3 × 3 × 2 × 2 ANOVA was calculated on the distinctiveness ratings assigned by the participants to each example behavior. Finally, two 3 (Preobservation Instructions) × 3 (Pilot Experience) between-subjects ANOVAs were calculated on the number of behavioral examples the pilots were able to list after watching the tape, both before and after receiving feedback regarding the targeted behaviors in the tape. The summaries for these ANOVAs are shown in Table 3. In the following we present the results by independent variable (main effect) and by interactions.

Behaviors Recognized and Described

Effects of experience. Experience had a significant main effect on the percentage of behaviors recognized by the pilots, $F(2, 50) = 5.07$, $p = .010$. Fisher LSD post hoc tests ($p < .05$) showed that the pilots in the ME ($M_{ME} = 53\%$) and HE ($M_{HE} = 52\%$) groups recognized significantly more of the targeted behaviors than did those in the LE group ($M_{LE} = 42\%$). There was also a significant effect of experience on the quality of behavior descriptions, $F(2, 50) = 3.61$, $p = .034$. Again, pilots in the two medium- and high-experience groups described the behaviors in more detail ($M_{ME} = 1.6$ and $M_{HE} = 1.6$, respectively) than did those in the low-experience group ($M_{LE} = 1.2$).

Effects of preobservation instructions. The preobservation instructions also had a significant effect on the percentage of behaviors that were recognized by the pilots, $F(2, 50) = 10.92$, $p < .001$. Fisher LSD post hoc tests ($p < .05$) indicated that pilots who had received both instructions to focus on the copilot and a list of behavior categories recognized significantly more of the targeted behaviors ($M_{OBS SPEC COPILOT} = 58\%$) than did pilots who had been told only to watch the copilot but had not been given the list ($M_{OBS COPILOT} = 47\%$). Pilots who had received both instructions to focus on the copilot and a list of behavior categories also recognized more of the targeted behaviors than did the pilots who had neither been told to...
### TABLE 3: Summary of Main Effects

#### Main Effect 1: Preobservation Instructions (between-subjects)

<table>
<thead>
<tr>
<th></th>
<th>Observe Crew (OBS CREW)</th>
<th>Observe Copilot (OBS COPILOT)</th>
<th>Observe Copilot (OBS-SPEC COPILOT)</th>
<th>Specific Behaviors</th>
<th>F (2, 50)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of behaviors recognized</td>
<td>42%</td>
<td>47%</td>
<td>58%</td>
<td>10.92</td>
<td>&lt; 0.001**</td>
<td></td>
</tr>
<tr>
<td>Quality of behavior descriptions&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.0</td>
<td>1.3</td>
<td>2.1</td>
<td>25.84</td>
<td>&lt; 0.001**</td>
<td></td>
</tr>
<tr>
<td>Distinctiveness ratings (retrospective)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5.5</td>
<td>5.4</td>
<td>5.7</td>
<td>0.57</td>
<td>0.569</td>
<td></td>
</tr>
<tr>
<td>Number of examples given before feedback</td>
<td>9.4</td>
<td>10.7</td>
<td>13.5</td>
<td>7.27</td>
<td>0.002**</td>
<td></td>
</tr>
</tbody>
</table>

#### Main Effect 2: Experience Level (between-subjects)

<table>
<thead>
<tr>
<th></th>
<th>Low (LE)</th>
<th>Medium (ME)</th>
<th>High (HE)</th>
<th>F (2, 50)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of behaviors recognized</td>
<td>42%</td>
<td>53%</td>
<td>52%</td>
<td>5.07</td>
<td>0.010*</td>
</tr>
<tr>
<td>Quality of behavior descriptions&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.2</td>
<td>1.6</td>
<td>1.6</td>
<td>3.61</td>
<td>0.034*</td>
</tr>
<tr>
<td>Distinctiveness ratings (retrospective)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5.6</td>
<td>5.9</td>
<td>5.3</td>
<td>3.82</td>
<td>0.029*</td>
</tr>
<tr>
<td>Number of examples given before feedback</td>
<td>9.8</td>
<td>11.8</td>
<td>12.0</td>
<td>1.96</td>
<td>0.150</td>
</tr>
</tbody>
</table>

#### Main Effect 3: Consequence Shown vs. Not Shown (within-subjects)

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>F (1, 50)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of behaviors recognized</td>
<td>27%</td>
<td>72%</td>
<td>342.98</td>
<td>&lt; 0.001**</td>
</tr>
<tr>
<td>Quality of behavior descriptions&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.8</td>
<td>2.1</td>
<td>25.84</td>
<td>&lt; 0.001**</td>
</tr>
<tr>
<td>Distinctiveness ratings (retrospective)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5.2</td>
<td>6.0</td>
<td>12.59</td>
<td>&lt; 0.001**</td>
</tr>
</tbody>
</table>

#### Main Effect 4: Positive vs. Negative Examples Shown (within-subjects)

<table>
<thead>
<tr>
<th></th>
<th>Negative</th>
<th>Positive</th>
<th>F (1, 50)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of behaviors recognized</td>
<td>55%</td>
<td>44%</td>
<td>16.54</td>
<td>&lt; 0.001**</td>
</tr>
<tr>
<td>Quality of behavior descriptions&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.5</td>
<td>1.5</td>
<td>0.01</td>
<td>0.964</td>
</tr>
<tr>
<td>Distinctiveness ratings (retrospective)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5.4</td>
<td>5.8</td>
<td>5.74</td>
<td>0.020*</td>
</tr>
</tbody>
</table>

N = 59, *p < .05, **p < .01

<sup>a</sup>Quality of behavior descriptions could range from 0 (no recognition) to 5 (example was recognized, described, and related to the underlying category of backup behaviors).

<sup>b</sup>Distinctiveness ratings could range from 1 (not at all) to 7 (very).
watch the copilot nor been given a list of behavior categories (\(M_{\text{OBS CREW}} = 42\%\)).

There was a similar significant main effect of preobservation instructions on the quality of the behavior descriptions given by the pilots, \(F(2, 50) = 25.84, p < .001\). Fisher LSD post hoc tests indicated that pilots who had received both information on whom to watch and also given a list of behavior categories to watch for described the modeled behaviors in better detail (\(M_{\text{OBS SPEC COPilot}} = 2.1\) ) than those who had just been told to watch the copilot (\(M_{\text{OBS COPilot}} = 1.3\) ) and those who had not been told only that they should focus on the copilot (\(M_{\text{OBS CREW}} = 1.0\) ).

**Interaction of experience and preobservation instructions.** Surprisingly, there was no significant interaction between condition and experience with respect to either recognition, \(F(4, 50) = 1.45, p = .250\), or the quality of the descriptions given by the pilots, \(F(4, 50) = 1.89, p = .127\). This indicated that, overall, all experience groups benefited about equally from receiving more detailed instructions.

**Effects of the salience cues.** Significant main effects were found with respect to the two salience cues employed in the videotape: whether a consequence was shown for a behavior or not, and regarding the quality (positive/negative) of each behavior. Overall, a higher percentage of behaviors were recognized when their consequences were shown than when they were not shown (\(M_{\text{CONS}} = 72\% \) vs. \(M_{\text{NOCONS}} = 27\%\)), \(F(1, 50) = 342.98, p < .001\). The quality of descriptions also changed significantly as a function of whether a consequence was shown or not, \(F(1, 50) = 96.29, p < .001\). Again, behaviors for which consequences were shown in the videotape were described in significantly more detail (\(M_{\text{CONS}} = 2.1\) ) than those for which no consequence was shown (\(M_{\text{NOCONS}} = 0.8\) ).

Further, proportionally more negative than positive examples of behavior were recognized (\(M_{\text{NEG}} = 55\%\) vs. \(M_{\text{POS}} = 44\%\)), \(F(1, 50) = 16.54, p < .001\). However, the quality of the descriptions for positive and negative examples did not differ, \(F(1, 50) = 0.001, p = .976\).

**Interactions of salience cues with preobservation instructions.** Whether or not a consequence was shown interacted with the preobservation instructions condition, \(F(2, 50) = 4.29, p = .019\). Showing the consequences improved performance of those pilots who had been told only to observe the crew in the instructions, raising their performance to the level of those whose attention was focused on the copilot (Figure 1).

Finally, the interaction between the type of the behaviors (positive/negative) and whether consequences were shown approached but did not reach statistical significance, \(F(1, 50) = 3.79, p = .053\). Inspection of the means showed that negative examples seemed to benefit more from showing their consequences than did positive examples. Likewise, the three-way interaction between preobservation instructions condition, type of behaviors, and whether their consequences were shown approached but did not reach statistical significance, \(F(2, 50) = 2.86, p = .058\).

**Pilots’ Rating of Behaviors Shown in the Videotape**

The ANOVA on the subjective ratings of distinctiveness that observers assigned retrospectively to each conflict showed a markedly different pattern of results than the previous analyses regarding recognition and description. Although there was a significant effect of experience on the ratings, \(F(2, 50) = 3.82, p = .029\), Fisher LSD post hoc tests (\(p < .05\) ) indicated that the only significant difference was between pilots in the highest-experience group (\(M_{\text{HE}} = 5.27\) ) and those in the medium-experience group (\(M_{\text{ME}} = 5.85\) ). There were no significant differences between pilots in the lowest-experience group (\(M_{\text{LE}} = 5.64\) ) and those of the other two groups. Further, there was no significant difference in distinctiveness scores as a function of preobservation instructions condition, \(F(2, 50) = 0.57, p = .569\).

There were significant differences in the distinctiveness ratings as a function of the behaviors’ characteristics: Behaviors whose consequences were shown were rated as more distinctive (\(M_{\text{CONS}} = 6.02\) ) than were those for which no consequences were shown (\(M_{\text{NOCONS}} = 5.15\) ), \(F(1, 50) = 12.59, p < .001\). Contrary to the results from recognition, however, positive examples were rated as *more* distinctive (\(M_{\text{POS}} = 5.79\) ) than negative examples (\(M_{\text{NEG}} = 5.38\) ), \(F(1, 50) = 5.74, p = .020\).
Positive Examples of Target Behaviors Generated

Before feedback. The number of positive examples that the pilots generated after the observation period, but before receiving feedback, was significantly related to both the number ($r = .51$, $N = 59$, $p < .001$) and the quality ($r = .49$, $N = 59$, $p < .001$) of recognized behaviors. A 3 instructions conditions $\times$ 3 experience levels ANOVA, however, showed a significant effect only of preobservation instructions condition, $F(2, 50) = 7.27$, $p = .002$. Fisher LSD post hoc tests ($p < .05$) indicated that pilots who had received both instructions only to observe the copilot and also had been given the list of target behaviors were able to generate significantly more examples of positive backup behaviors ($M_{OBS\ SPEC\ COPilot} = 13.5$) than were those who had been told only to observe the copilot ($M_{OBS\ COPilot} = 10.7$) and also than those who had been given only instructions to record good and bad examples of crew behaviors ($M_{OBS\ CREW} = 9.4$).

After feedback. After the pilots were shown the list of the targeted behaviors in the videotape, correlations between the number of generated examples and the number and quality of recognized behaviors were found to be not statistically significant ($r = .17$, $N = 59$, $p = .200$ and $r = .06$, $N = 59$, $p = .645$, respectively) and significantly lower than before ($z = -2.93$ and $z = -3.53$, respectively). Also, an ANOVA showed no significant effects of either instructions condition or experience on the number of examples generated after feedback.

Summary

Both experience and preobservation instructions were associated with significant differences in the percentage of targeted behaviors recognized and with respect to the quality of the descriptions trainees could give. In each case, pilots who had been given instructions to watch for specific classes of copilot behaviors did better than those who were told only to observe the copilot and those who were merely told to observe the crew. Further, pilots of medium and high experience in this study did better than those in the low-experience group. The type of behavior being shown (positive vs. negative examples) and whether or not a consequence was shown significantly affected the percentage of targeted behaviors that were recognized and/or the quality of the descriptions given by the pilots. Ratings of behavior distinctiveness and the number of behaviors pilots were able to generate (list) showed very different patterns of results.

![Figure 1](image_url). Interaction between preobservation instructions and whether or not consequences were shown of the examples, as measured by the numbers of behaviors recognized by the observers.
DISCUSSION

Observational learning is one of the most popular training approaches. It is appealing because it is cost-effective and is based on the strength of decades of research supporting the utility of behavioral observation as a training technique. The present study sought to investigate certain factors that were hypothesized to influence the utility of this intervention for one complex application for which it is currently being used: the training of coordination skills required by first officers in commercial aviation.

A troublesome initial finding was the low percentage of behaviors recognized by participants. The behaviors were selected and performed in such a manner that they were consistent with established practice and salient in the videotape. Nevertheless, participants recognized only about 49% of the targeted behaviors on average. Even participants who had been specifically told to observe the copilot and to look for specific examples of behaviors in six categories recognized, on average, only 58% of the targeted behaviors. This is consistent with similar research on error detection by pilots, which, using a similar methodology as the one employed here, showed error detection rates from videotapes of around 30% to 40%, depending on observer experience (Doireau, Wioland, & Amalberti, 1997).

Together, findings from the study by Doireau et al. (1997) and from our current study highlight the need to understand the factors which influence cue salience in these types of videos so that all of the targeted training events are perceived by the subjects. The low rate of recognition also has troublesome implications regarding one of the most prominent behavior classes in team work, alternately referred to as “backup behaviors,” “compensatory behaviors,” or “monitoring and challenging.” The National Transportation Safety Board (1994) has frequently cited “lack of monitoring and challenging” as one of the main contributing factors in transportation accidents. The current data suggest that even under relatively low workload conditions, pilots have trouble recognizing mistakes that others make. This is in line with other research (e.g., Jentsch, Bowers, Martin, Barnett, & Prince, 1998) suggesting that junior pilots often fail to monitor and challenge, not because they are not assertive enough but also because they do not recognize a situation as requiring action. Thus one question that needs to be addressed is how pilots can improve their ability to recognize positive and negative behaviors, both in training and in their work environment.

The results also indicated that novices are at particular risk for missing key stimuli embedded in training videos. In fact, on average, novices observed less than half (42%) of the stimuli that were used as eliciting events in this training video, and only about half (54%) did so when they were given specific instructions on whom to watch and what to look for. This is problematic for training developers because novices are the typical audience for this training intervention but appear to be the least able to benefit from it. Consequently, there is a need to identify those factors that increase cue salience, especially for the novice group.

Conversely, the results also indicated that more experience is not always a guarantee of better performance. We found no significant differences in recognition and description of targeted behaviors between the medium- and high-experience groups. This suggests that more experience is beneficial only to a point; beyond that point, the data reported here support the contention that it is the type and variety of experience that builds expertise, not merely the amount.

The results of this study also offer some directions for increasing the degree to which targeted behaviors will be recognized. Perhaps most important, it appears that the nature of preobservation instructions given to participants can have an important effect on the degree to which specific behaviors are recognized. In this study participants who simply watched the tape after being told to observe the crew’s behaviors recognized significantly fewer of the targeted behaviors than did those who were given more explicit directions. This was true of both instructions related to the actor (who to watch) and the behaviors to be observed (what to watch for). Furthermore, the quality of the observation was also better in the latter condition.

Although these findings are not surprising, the lack of an interaction between expertise
and preobservation instructions across all experience conditions was. Given that observational learning can be hypothesized to require engagement in cognitive processes similar to those in practice/experience, a case could certainly have been made that novices would benefit more from specific instructions than would experts. The absence of the interaction, however, suggests two things. First, all pilots can benefit from specific instructions of what to look for in training videos. Second, as experience increases, pilots may be able to do with less specific instructions (as indicated by the fact that high-experience pilots needed to be told only to watch the copilot and did not significantly improve after that, whereas medium-experience pilots did better when they were not only told to watch the copilot but were also given the list of six specific behavioral categories to look for). However, this should not be taken to mean that more experienced pilots do not need to be prepared before watching modeling displays. Clearly, even pilots in the high-experience group had to be told at least whom to watch and for what general behaviors; merely telling them to watch the tape was not enough.

The selection of targeted events can also influence the degree to which targeted behaviors are observed. Negative events were recognized more often, an effect that has been reported in the past. It has been suggested that the effectiveness of using negative behaviors in modeling interventions might be attributable to increasing variability of the stimulus set (Bruner, Goodnow, & Austin, 1956; Reed, 1987) or to “unfreezing” inappropriate attitudes regarding unacceptable behaviors (Lewin, 1951). However, the present study suggests that the improved efficacy of these behaviors might simply be a result of greater recognizability of these behaviors. Including the consequences of targeted events appears to be another manner in which the salience of target behaviors can be increased.

Finally, it is important to recognize that pilot ratings of distinctiveness (i.e., salience) showed a completely different pattern of results than did the actual recognition data. This suggests that training designers should not merely rely on ratings of distinctiveness but should also collect recognition data when creating videotapes for training using observational learning.

Future Research

In future research, recognition of targeted behaviors needs to be linked to measures of behavioral change, be it in the form of reproduction (“can do”) measures or generalization/transfer (“will do”) measures on the job. As the current investigation shows, there appears to be a correlation between recognition and a simple form of generalization. At the same time, however, the ability to generate examples of appropriate behaviors may not be entirely predictive of the ability (and disposition) to perform them on the job. It is in these areas that future investigations should study the generalizability of the current results.

ACKNOWLEDGMENTS

We thank the students and flight instructors at Comair Aviation Academy who participated in this research, as well as Kim Smith-Jentsch for her comments on this manuscript. This research was supported in part by a collaborative agreement between the Federal Aviation Administration, Office of the Chief Scientific Advisor for Human Factors, AAR-100, and the Naval Air Warfare Center Training Systems Division. All opinions stated in this paper are those of the authors and do not necessarily reflect the opinion of the Federal Aviation Administration, the U.S. Navy, or the University of Central Florida, unless otherwise designated by official document. An earlier version of this research was presented at the 1997 Annual Meeting of the Society for Industrial/Organizational Psychology (SIOP), St. Louis, MO.

REFERENCES


Florian Jentsch is an associate scientist in the Department of Psychology at the University of Central Florida. He received his Ph.D. in human factors psychology from the University of Central Florida in 1997.

Clint Bowers is an associate professor in the Department of Psychology at the University of Central Florida. He received his Ph.D. in clinical/community psychology from the University of South Florida in 1987.

Eduardo Salas is a professor in the Department of Psychology and Institute for Simulation and Training at the University of Central Florida. He received his Ph.D. in industrial/organizational psychology from Old Dominion University in 1984.

Date received: November 2, 1999

Date accepted: January 8, 2001