Configural Display Design Techniques Considered at Multiple Levels of Evaluation

Kevin B. Bennett and **Brett Walters**, Wright State University, Dayton, Ohio

Two studies were conducted to examine issues in the design and evaluation of configural displays. Four design techniques (bar graphs/extenders, scale markers/scale grids, color coding/color layering/color separation, and annotation with digital values) were applied, alone and in combination, to a baseline configural display, forming 10 displays. Two qualitatively different evaluations assessed performance for (A) low-level data probes (quantitative estimates of individual variables) and (B) system control and fault detection tasks. Three of the four design techniques improved performance significantly for low-level data probes (color coding was the exception). A display with digital values only (i.e., no analog configural display) produced the poorest performance for control/fault detection tasks. When both levels of evaluation are considered, a composite display (configural display with all four techniques applied) was clearly the most effective. Overall, the findings obtained in the two experiments provide very limited evidence for the generalization of results between evaluations. The two levels of evaluation, the display manipulations, and the patterns of results are considered in terms of a cognitive systems engineering evaluation framework. General implications for the evaluation of displays and interfaces are discussed. Actual or potential applications include design techniques to improve graphical displays and methodological insights to focus and improve evaluation efforts.

INTRODUCTION

An ongoing research program has explored issues in the design and evaluation of configural displays (Bennett & Flach, 1992; Bennett, Nagy, & Flach, 1997; Bennett, Payne, Calcaterra, & Nittoli, 2000; Bennett, Toms, & Woods, 1993). This type of display maps several individual variables into a single geometrical form. Changes in the individual variables cause the overall pattern, or configuration, of that form to vary. One focus of these research efforts has been on fundamental issues in the design of these displays. These include issues in perception and pattern recognition as well as the quality of the specific mappings between graphical representations and domain semantics. A second focus has been on fundamental issues in evaluation, including the use of multiple methodologies. Factors in both design and evaluation were investigated in the present study: Alternative versions of a configural display were evaluated using two qualitatively different methodologies. We begin with a description of the basic configural display and the design techniques that were evaluated.

The basic configural display included four variables: two variables that were plotted on the *y* axis and two variables that were plotted on the *x* axis. A single geometrical form (a rectangle) was plotted at the intersection of these four variables. This rectangle could change in size, shape, and location within the *x*-*y* grid (for a more complete description of the display, see Bennett et al., 1993).

Ten displays were implemented using four design techniques (the scales, color, bar-ex, and digital techniques). The *baseline* configural

Address correspondence to Kevin B. Bennett, Department of Psychology, 335 Fawcett Hall, Wright State University, Dayton, OH 45435; kevin.bennett@wright.edu. **HUMAN FACTORS,** Vol. 43, No. 3, Fall 2001, pp. 415–434.

display (Figure 1) had no techniques applied. The scales design technique added scale markers and scale gridlines to the baseline (Figure 2). The color design technique (Figure 3) used chromatic and luminance contrast to add color coding, visual layering, and visual separation to the baseline display. The bar-ex design technique (Figure 4) incorporated bar graphs for each individual variable and "extenders" that connected them to the configural form. Eight of the 10 displays were formed through a factorial combination of these three design techniques, applied at two levels (present or absent). The final two displays incorporated the fourth design technique – digital values. The *composite* display had the scales, color, and bar-ex design techniques applied and was also annotated with digital values (Figure 5). The *digital* display consisted of digital values alone (no analog configural display, Figure 6).

The impact of these display manipulations on performance was assessed using alternative methodologies falling into two distinct categories. These categories will be considered in terms of the cognitive systems engineering evaluation framework outlined by Rasmussen, Pejtersen, and Goodstein (1994). This framework consists of five levels of evaluation, ranging from highly controlled laboratory research to field studies. The levels are differentiated by *constraint envelopes* or *constraint boundaries.* The overall constraint envelope at a particular level of evaluation is a joint function of three independent, but interacting, sources of constraints: task, interface, and observer.

One set of constraints is associated with the task(s) to be performed: Each particular task or set of tasks will introduce a set of cognitive demands that must be met. The functionality/design of the interface introduces another set of constraints: Particular characteristics of the interface will introduce cognitive demands that will vary in terms of the nature and amount of cognitive resources that are required. The final constraints are those introduced by an individual's cognition/perception/action capabilities and limitations. Overall levels of performance will be determined by the extent to which the three sets of constraints are well matched. The evaluations reported here are Boundary 1 and 3 evaluations. The general characteristics of each of these boundaries will be described in greater detail.

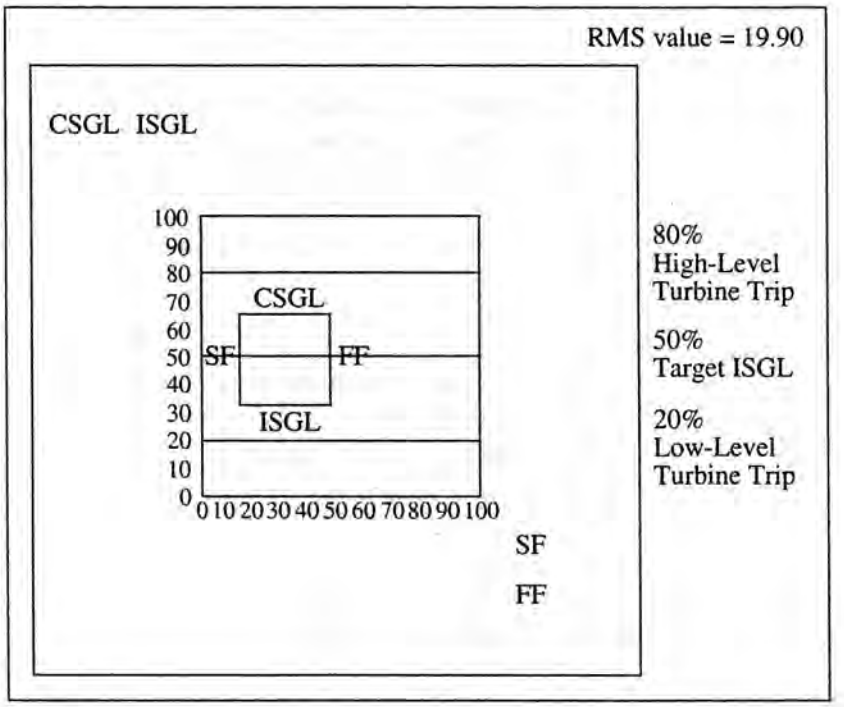

Figure 1. The baseline configural display.

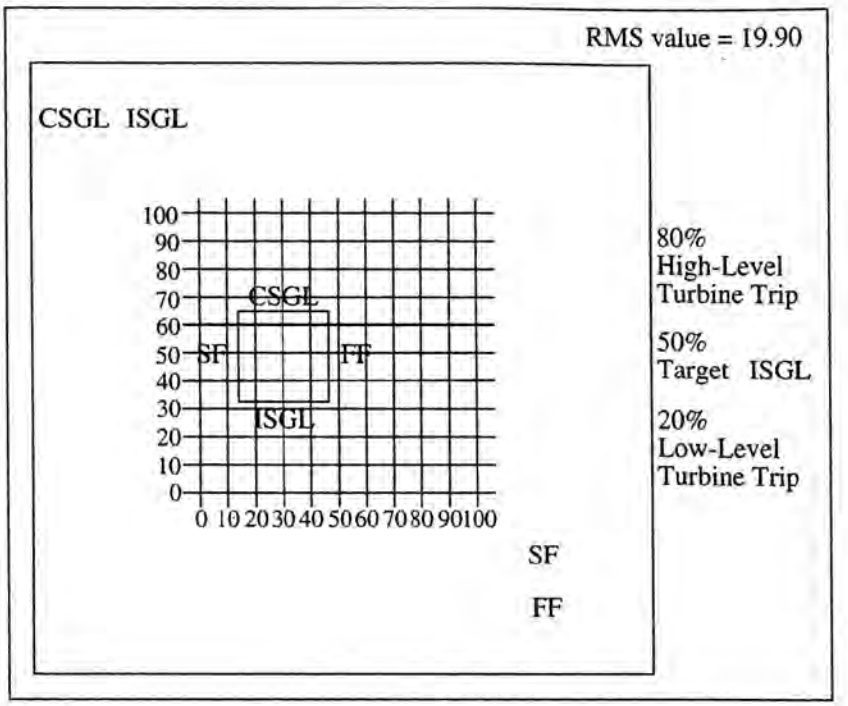

Figure 2. The baseline plus the scale markers/gridlines (scales) design technique.

Boundary Level 1 (Controlled Mental Processes)

Rasmussen et al. (1994) describe experiments conducted at Boundary 1 as "the evaluation of actor-related issues in an environment that corresponds most closely to the traditions of experimental psychology" (p. 205). Generally speaking, a key characteristic is tight experimental control: The experimental instructions are explicit, the criteria for effective performance are well defined, and the number of alternative strategies that could be developed to complete the task are relatively limited.

One type of display design research conducted at Boundary 1 has the primary goal of assessing the relationship between human and display. The task constraints are often simple; for example, "Report the value of low-level data (individual variable) or high-level properties (relationship between variables)." These experimental tasks are essentially defined in terms of the display itself. Therefore, the display constraints are determined by the characteristics of the visual features that were chosen. Although the visual features may have been designed with a particular domain in mind, the evaluation itself is relatively domain-independent (i.e., the characteristics of the underlying domain would presumably have little impact on the performance of these types of tasks). The observer constraints arise from basic cognition/perception/action capabilities. Performance is usually measured in domain-independent terms such as accuracy and latency of response; the level of performance will reflect the extent to which the observers' perceptual systems are sensitive to variations in the visual features (i.e., how well observers can obtain the information that has been encoded into the display).

Boundary Level 3 (Controlled Task Situation)

The overall constraint envelope imposed by a Boundary 3 evaluation is different from that of a Boundary 1 evaluation. The differences are manifest in all three sources of constraints. The experimental test bed simulates some portion of a real work domain. Thus the tasks to be completed are defined in terms of an underlying domain (e.g., system control), rather than in terms of the display itself. Consequently, these tasks tend to be considerably more complex.

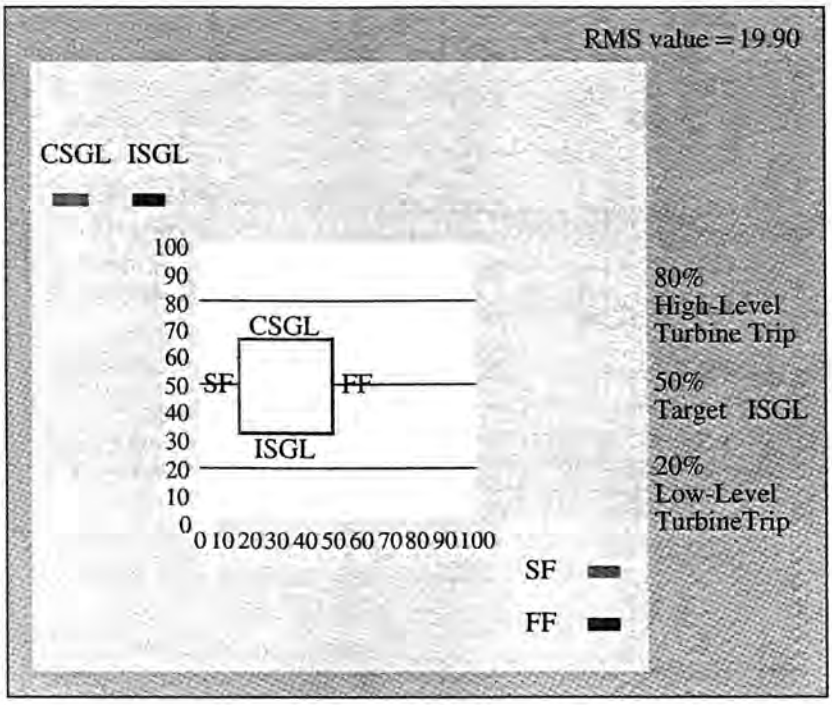

Figure 3. Baseline plus the color/layering/separation (color) design technique.

The metrics used to gauge performance are also usually defined in terms of the system (e.g., how many completed parts are produced in a manufacturing plant). The observer constraints are expanded considerably: Domain-specific skills and knowledge that are related to the goals, constraints, and physical/functional characteristics of the system become important at this level. The experimental instructions and performance criteria are necessarily less explicit, and the potential for alternative strategies (both effective and ineffective) is considerably greater. At this level of evaluation, the display introduces a set of constraints that are also dependent on the quality of the mapping between the domain semantics (relationships, properties, goals, constraints) and the corresponding visual features.

Boundary 1 and 3 Evaluations in the Present Experiment

The Boundary 3 evaluation was conducted in the domain of process control. A part-task simulation was used that replicated the basic dynamic characteristics of a single nuclear power plant steam generator during start-up. Briefly, the manual control of the feedwater task involves the control of mass flowing into (the rate of feedwater flow, called *feed flow*) and out of (the rate of steam flow, called *steam flow*) a steam generator so that the level of coolant inside (indicated steam generator level, or *indicated level*) is maintained between upper and lower limits. Control is complicated by the fact that energy inflow/outflow also influences indicated level but at a different time constant (i.e., by producing counterintuitive shrink/swell effects). To assist in control, a quickened variable was developed that provides an estimate of indicated level that is not confounded by these energy effects: compensated steam generator level (*compensated level*).

The participants performed a control task in the Boundary 3 evaluation: They changed the rate of feed flow to (A) bring indicated level to a goal value (50%) as quickly as possible and (B) maintain indicated level as close to this goal as possible. In addition, participants were provided with information about two types of faults (either a steam generator leak or a stuck valve) and were instructed to report the occurrence of a fault. Thus this evaluation focused on the extent to which the various design techniques supported the participants in the per-

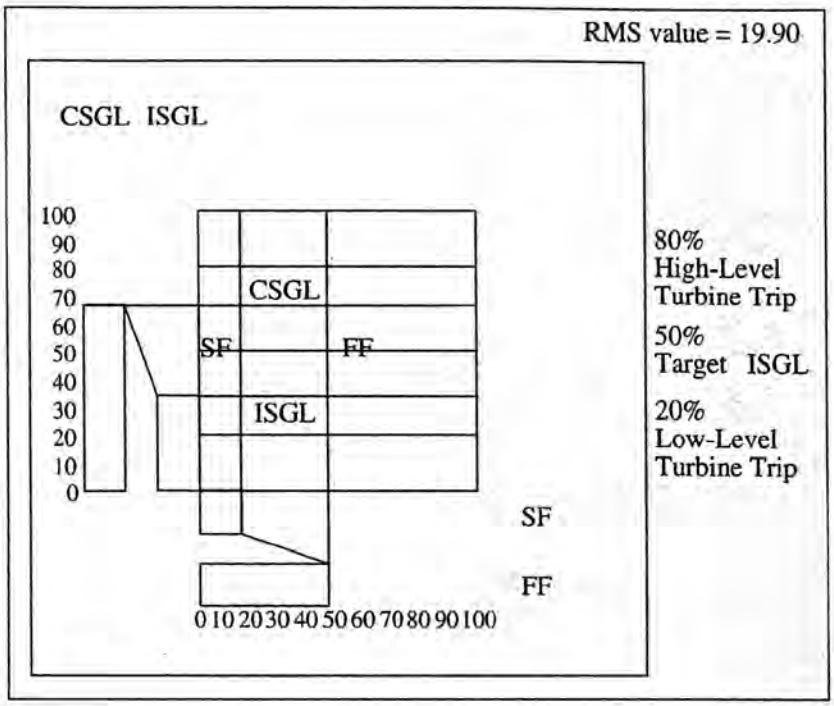

Figure 4. Baseline plus the bar-graphs/extenders (bar-ex) design technique.

formance of relatively complex process control and monitoring tasks.

The Boundary 1 evaluation focused on more basic issues in the design of configural displays. Participants performed low-level data probes that required them to provide quantitative estimates of one of the four individual variables. Supporting performance at these types of tasks (*low-level data*) has been one concern in the design of configural displays (Bennett & Flach, 1992; Wickens & Carswell, 1995). Bennett et al. (2000) demonstrated that configural displays could be designed to provide better support for these tasks: Performance for a display with multiple design techniques applied (similar to Figure 5) was significantly faster and more accurate than the same configural display (similar to Figure 1) without the applied techniques. Thus this evaluation focused on the relative contributions of the various design techniques to improved performance for the extraction of low-level data.

EXPERIMENT 1

Method

Participants. Eight students (five men and three women) participated in the experiment and were paid $5.00/hr. The participants' ages ranged from 20 to 32 years. They had normal or normal-corrected vision with no color-blindness deficiencies. Three of the participants had completed similar experiments.

Apparatus. All experimental events were controlled by a general-purpose laboratory computer (Sun Microsystems, Inc., 4-110 Workstation, Palo Alto, CA) located in an enclosed experimental room. A color video monitor (40.64 cm, 1152 × 900 resolution) and a standard keyboard were used.

Simulation model. For a more detailed description of the simulation model, see Bennett et al. (1993).

Stimuli. In total, 10 displays were evaluated. The baseline display (Figure 1) mapped four variables into a rectangle appearing inside a display plotting matte (corresponding to the *x* and *y* axes, which measured 10.16 cm × 10.16 cm and subtended a maximum visual angle of 11.49° both horizontally and vertically, assuming a participant seated 50 cm away). A display background matte measuring 17.78 cm high and 24.46 cm wide enclosed the display. All graphical elements were black; the background was dark gray.

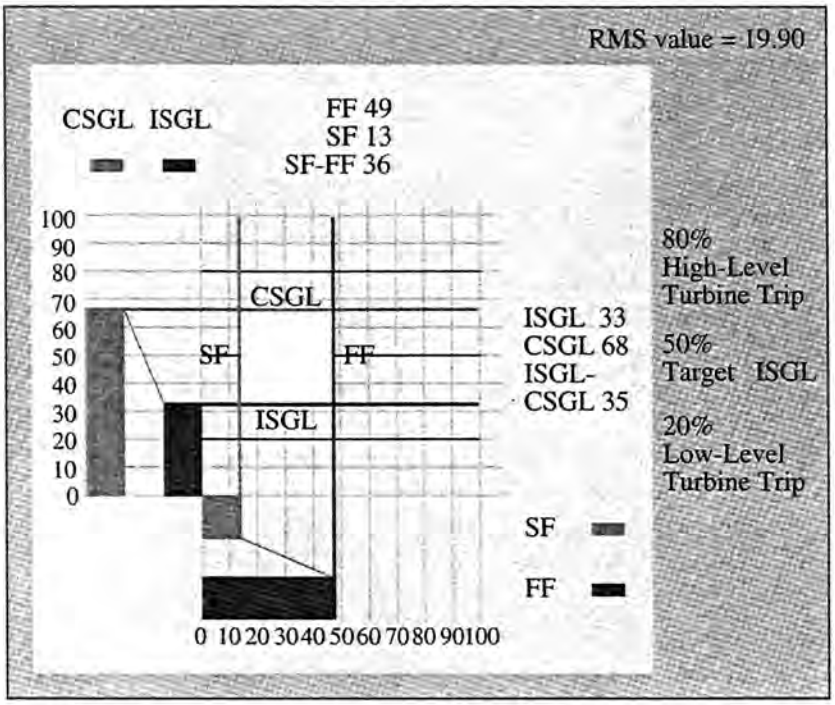

Figure 5. Composite display: the baseline with the scales, color, bar-ex, and digital design techniques.

When possible, the remaining displays used the same sizing and labeling conventions as the baseline display. The scales design technique included (A) scale markers next to the labels on the *x* and *y* axes and (B) scale gridlines in the plotting matte (Figure 2). The color technique (Figure 3) used chromatic and luminance contrast to stratify categories of information in the display (background matte – medium gray; plotting matte – light gray; rectangle – off-white; color codes for individual variables – green, purple, blue, mustard; trip set points – dull red; target value for indicated level – white). The bar-ex design technique (Figure 4) included bar graphs (one-pixel black outlines that were 1.12 cm wide) and extenders (connecting the bar graphs to the configural display). The display axes were displaced downward (*x* axis) or to the left (*y* axis) to accommodate the bar graphs.

Four additional displays that used multiple techniques are not represented in the figures. The two remaining displays contained digital values (individual variables and two relationships between variables). The composite display (Figure 5) applied all four design techniques. The digital display (Figure 6) had digital values only, without the analog configural display. All displays were updated with information from the simulation model every 2 s, and random noise ranging from −1.5% to +1.5% was added to the value of each variable displayed (this noise did not change the values in the mathematical model).

Procedure. Each participant completed an introductory session (1.5 hr), three practice sessions (1 hr each), and eight experimental sessions (1 hr each). In the introductory session all participants were given a complete description of the system (graphic representations and verbal descriptions of components, variables, causal relationships, and faults). No discussion of specific control strategies was provided during the course of the experiment.

The participants were tested individually in an enclosed room. Each trial lasted 4 min. Steam flow and feed flow were 0% initially; indicated level and compensated level were 35% initially. Every 2 s, a 1% increase could occur in steam flow (50% probability), as long as steam flow was less than 20%. Participants changed feed flow by pointing and clicking on one of four boxes (increasing or decreasing feed flow by 1% or 4%). They were instructed

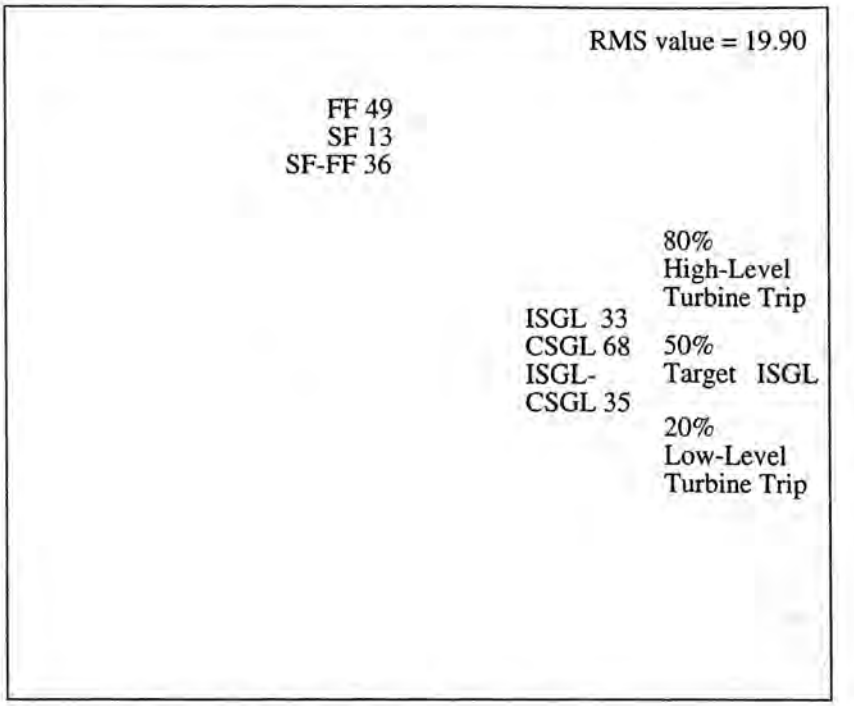

Figure 6. Digital display: digital design technique alone (without analog configural display).

to provide control inputs that moved the indicated level to a target level (50%) quickly, to maintain the indicated level close to this target level, and to avoid crossing setpoint boundaries. Continuously updated root mean square (RMS) error scores and auditory feedback for the indicated level boundary crossings (four tones) were provided.

Two faults could occur. One fault simulated a steam generator leak by decreasing the value of the indicated level by 0.33% at 2-s intervals. The second fault simulated a stuck valve: Control input to feed flow caused the representations on the screen to change (commanded value) but had no impact on the simulation (actual value). When a fault was present, it began between 30 and 90 s into a trial. The participants were instructed to detect faults as accurately and quickly as possible and to indicate the presence of a fault by pointing and clicking on a dedicated box. Feedback on the presence or absence of a fault was provided after each trial. Each fault occurred once for each display during the course of the experiment; combinations of fault and display were counterbalanced across participant and day. The presentation order for the 10 displays in a session was random. Fault trials occurred within the first 10 trials in a session; additional, nonfault trials for those displays were readministered at the end of the session. Thus 12 or 13 trials were completed in a session (with an average of 2.5 fault trials).

Four low-level data probes (one for each individual variable) were completed during an experimental trial. They occurred in four time windows (25–55 s, 65–95 s, 105–135 s, and 145–175 s) and were administered when the next screen update was scheduled to occur. An auditory tone sounded, a description of the probe was presented (e.g., "Enter % value for Steam Flow"), and participants entered a numeric value via the keyboard. The display remained visible at all times. The participants were instructed to respond to probes as accurately and quickly as possible. Feedback on both accuracy and latency was provided. A probe was readministered if the observer entered a value that was outside the acceptable range (0–100) or if the observer changed his or her estimate before completing a probe (e.g., deleting an original estimate and entering a new one). The color coding for the four variables was counterbalanced across participants.

RESULTS

A similar procedure was followed for the majority of analyses. Outliers were identified using the test described in Lovie (1986, pp. 55–56):

$$T_1 = (x_{(n)} - \bar{x}) / s, \quad (1)$$

in which $x_{(n)}$ is a particular observation (one of n observations), $\bar{x}$ is the mean of those observations, and s is the standard deviation of those observations. Nonparametric tests were conducted to determine if the outlier distribution was random (none was significant). The statistical analyses performed were a set of 11 preplanned contrasts. Table 1 provides a numeric label for each contrast, a verbal description of the contrast, and the displays with the associated contrast weights. The family-wise error rate was controlled using the modified Bonferroni test (Keppel, 1982), with an adjusted significance level of 0.041. Additional tests for simple main effects were conducted when an interaction contrast was significant.

Boundary 3: Control Performance

Six measures of control performance were considered. Acquisition time was measured from trial initiation until the indicated level first crossed into a target band (45% to 55%). Settling time was measured from trial initiation until the indicated level crossed and remained inside the band for the remainder of the trial. Four estimates of control error (Poulton, 1974) were considered during a final tracking phase (starting at the average settling time across all participants and ending at trial completion). The formula for RMS was

$$\sqrt{\Sigma(X - 50)^2 / N}, \quad (2)$$

in which X is the indicated level for an update and N is the number of updates. The formula for constant position error is

$$\Sigma(X - 50) / N. \quad (3)$$

The formula for modulus mean error is

$$\Sigma abs(X - 50) / N. \quad (4)$$

The formula for standard deviation of the error is

$$\sqrt{\Sigma(X - \bar{X})^2 / N - 1}, \quad (5)$$

in which $\bar{X}$ is the mean value of the indicated level across updates.

Preliminary analyses revealed two distinct groupings of performance at these tasks; three of the eight participants were unable to control the system effectively. Because of the substantial qualitative and quantitative differences in performance, their data were not considered in the Boundary 3 analyses. These differences are described in greater detail in the "General Discussion."

Outliers were identified in the nonfault trials for acquisition (five scores, 1.25%), RMS error (four scores, 1.00%), constant position error (two scores, 0.50%), modulus mean error (four scores, 1.00%), and standard deviation of the error (five scores, 1.25%) measures. The preplanned contrasts were conducted for the six control measures outlined previously (both nonfault and reservoir leak fault trials). The significant contrasts are listed on the right side of Table 2; the means for each display are illustrated in Figure 7a (nonfault trials).

Boundary 3: Fault Detection

The ability to discriminate between fault and nonfault trials was assessed using signal detection measures. False alarm rates were calculated for nonfault trials, and the planned comparisons were conducted. The significant contrasts are listed on the right side of Table 2. Hit rates were calculated for each of the two faults, and the planned comparisons were conducted on the combined data set. No comparisons were significant. The latency of fault detection was measured from fault onset until a participant response or the end of a trial (a cutoff corresponding to the largest time window common to all trials, 152 s, was applied). No contrasts were significant.

Boundary 1: Low-level Data

Accuracy (error magnitude) was measured by computing the absolute value of the difference

TABLE 1: Displays and Contrasts

Verbal description	Display									
	Baseline	Scales	Bar-ex	Scales, Bar-ex	Color	Color, Scales	Color, Bar-ex	Color, scales, Bar-ex	Composite	Digital
1. Main effect – scales	1	–1	1	–1	1	–1	1	–1	0	0
2. Main effect – color	1	1	1	1	–1	–1	–1	–1	0	0
3. Main effect – bar-ex	1	1	–1	–1	1	1	–1	–1	0	0
4. Interaction effect – scales & color	1	–1	1	–1	–1	1	–1	1	0	0
5. Interaction effect – scales & bar-ex	1	–1	–1	1	1	–1	–1	1	0	0
6. Interaction effect – color & bar-ex	1	1	–1	–1	–1	–1	1	1	0	0
7. Interaction effect – color, scales & bar-ex	1	–1	–1	1	–1	1	1	–1	0	0
8. Digital values vs. no digital values	–1	–1	–1	–1	–1	–1	–1	–1	4	4
9. Digital display vs. composite display	0	0	0	0	0	0	0	0	–1	1
10. Digital display vs. all others	–1	–1	–1	–1	–1	–1	–1	–1	–1	9
11. Composite display vs. all others	–1	–1	–1	–1	–1	–1	–1	–1	9	–1

Note: Numerical codes before verbal descriptions indicate contrast numbers.

between the participant's estimate of a variable and the actual value as it appeared on the screen (i.e., variable plus noise). Response time was measured from the appearance of the prompt until the first digit of the participant's response (1/100 s accuracy). Of the 3200 probes that were administered, 129 were identified as either accuracy or latency outliers (4.03%). The significant preplanned comparisons for both accuracy and latency are listed in the left side of Table 2; the means for each display are illustrated in Figure 7b.

DISCUSSION

The results of Experiment 1 indicate that performance at Boundary 3 (system control and fault detection) varied as a function of the design techniques that were applied. A primary finding was that the display with digital values only (digital) did not support system control effectively (see Figure 7a). Contrast 9 revealed that significantly degraded control performance was obtained for the digital display (relative to the composite display) during both fault trials

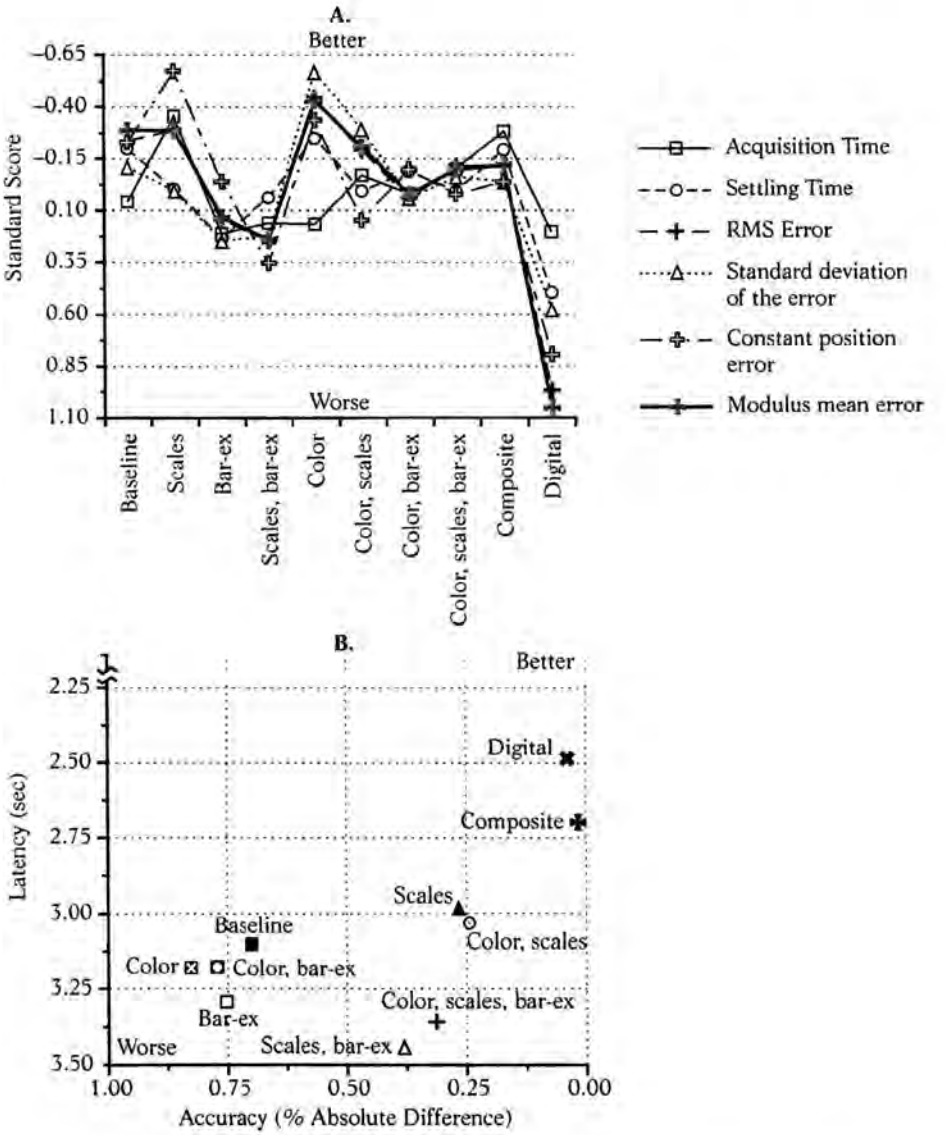

Figure 7. Results obtained for displays in Experiment 1. (a) Boundary 3 evaluation: average performance (standard scores) for the six measures of system control during nonfault trials. (b) Boundary 1 evaluation: latency (in seconds) and accuracy (error magnitude) for the low-level data probes.

TABLE 2: Significant Contrasts in Experiment 1

Contrast number	Boundary 1*		Boundary 3**
	Accuracy	Latency	Dependent Measures
1.	0.0001[a]		FAR 0.02[a]
3.		0.002	
5.		0.0003	
8.	0.0001	0.002	
9.		0.006[b]	RMS fault 0.0005[bc] CP fault 0.04[bc] MM fault 0.02[bc] MM 0.05[bc]
10.	0.0002	0.0008	
11.	0.0001[ac]	0.008[ac]	FAR 0.03[ac]

Note: FAR = false alarm rate; RMS = root mean square error; CP = constant position error; MM = modulus mean error. A "fault" designation indicates that the significant contrast was obtained for fault trials.
[a] Contrasts reveal the same pattern of results between boundaries.
[b] Contrasts reveal a different pattern of results between boundaries.
[c] Contrasts indicate the composite display improved performance at both boundaries.
* $F(1, 7)$, $p <$.
** $F(1, 4)$, $p <$.

(RMS error, constant position error, and modulus mean error) and nonfault trials (modulus mean error). Two additional findings were that the composite display produced significantly better false alarm rates, compared with all other displays (Contrast 11), and that the presence of scales/gridlines also produced significantly better false alarm rates (Contrast 1).

The design techniques also influenced performance significantly at the Boundary 1 evaluation (low-level data probes). A primary finding was that the presence of digital values improved performance dramatically. All contrasts comparing the two displays with digital values to other displays (Contrasts 8, 10, and 11 for both accuracy and latency) were significant (see Table 2). This is readily apparent in Figure 7b, which illustrates the clear separation of the digital and composite displays from all other displays. Contrast 9 for latency revealed that the digital display was the best display for the Boundary 1 evaluation, producing significantly lower response times than the composite display. The scales design technique also improved performance for low-level data probes significantly (Contrast 1 for accuracy).

The remaining two design techniques had either no effect or a negative effect on performance. The color (color coding/layering/separation) design technique neither facilitated nor degraded performance significantly. The bar-ex (bars/extenders) design technique degraded the latency for low-level data probes significantly (Contrast 3). The interaction between the bar-ex and the scales techniques (Contrast 5) indicated that the increases in latency associated with the bar-ex technique were significant when the scales technique was applied at the same time.

Consideration of the pattern of results between boundaries reveals several insights. There were three contrasts (Contrasts 1, 9, and 11) in which a display manipulation produced significant contrasts at both boundaries. For Contrast 9, an opposite pattern of results was obtained at the two boundaries: The digital display improved performance at Boundary 1 (low-level data probes) but degraded performance at Boundary 3 (system control during reservoir leak fault and nonfault trials). These results are shown in Table 2.

The remaining two contrasts (1 and 11) revealed a similar pattern of results between boundaries. The composite display and the scales technique produced improvements in false alarm rates. In addition, the composite

display was effective in supporting performance at both boundary levels. The contrasts supporting this observation are labeled with a "C" superscript in Table 2. A second experiment was conducted, with only a few aspects of the methodology and stimuli changed.

EXPERIMENT 2

Method

The apparatus, simulation model, displays, procedure, and participants were identical to those in Experiment 1, with a few exceptions. One participant (a woman with poor control performance) could not continue because of personal reasons. She was replaced by another woman participant, who completed an introductory session and five experimental sessions of practice before participating in the experiment. All participants completed eight experimental sessions. In addition, the bar-graph/extender technique was redesigned (see Figure 8). The numeric labels remained adjacent to the display grid, as in all other conditions, and the bar graphs/extenders were shifted either to the left or down, relative to their position in Experiment 1 (compare Figures 5 and 8). In addition, the lines connecting the pairs of bar graphs that were present in Experiment 1 were removed. In Experiment 1 steam flow ramped gradually up to 20%; In Experiment 2 steam flow ramped gradually up to 20%, but also oscillated (could decrease as well as increase) as a function of three sine waves. The rate of a reservoir leak fault was also decreased from 0.33% to 0.25%.

RESULTS

Boundary 3: Control Performance

The two participants who were unable to control the system effectively in Experiment 1 were again unable to do so in Experiment 2. Their data were not considered in the Boundary 3 analyses. Outliers were identified in the acquisition (11 scores, 2.29%), RMS error (6 scores, 1.25%), constant position error (3 scores, 0.63%), modulus mean error (8 scores, 1.67%), and standard deviation of the error (8 scores, 1.67%). The significant contrasts are listed on the right side of Table 3; the display means for all six measures during nonfault trials are illustrated in Figure 9a.

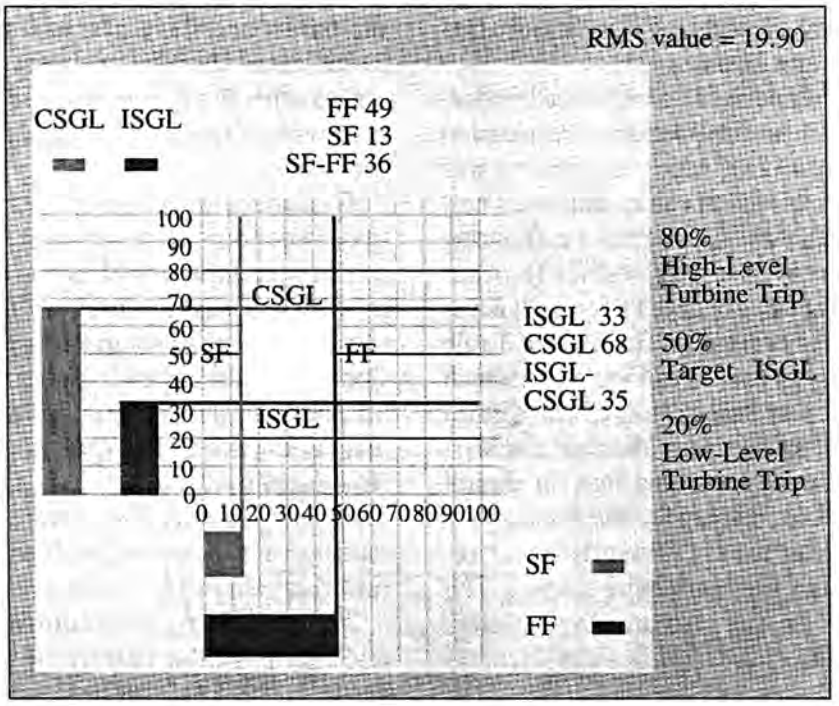

Figure 8. Composite display used in Experiment 2, including the revised bar-ex design technique.

TABLE 3: Significant Contrasts in Experiment 2

Contrast number	Boundary 1*		Boundary 3**
	Accuracy	Latency	Dependent Measures
1	0.0001		
3	0.005		
4	0.04		
5	0.002		
6	0.03	0.04	
8	0.0001	0.0003	
9		0.0002[b]	AT 0.03[bc]
10	0.0001[b]	0.0001[b]	SD 0.04[b]
11	0.0001[ac]	0.02[ac]	AT 0.03[ac] FAR 0.001[ac]

Note: AT = acquisition time; SD = standard deviation of the error; FAR = false alarm rate.
[a] Contrasts reveal the same pattern of results between boundaries.
[b] Contrasts reveal a different pattern of results between boundaries.
[c] Contrasts indicate the composite display improved performance at both boundaries.
* $F(1, 7)$, $p <$.
** $F(1, 5)$, $p <$.

Boundary 3: Fault Detection

Analyses for hit rate, false alarm rate, and detection latency for the two faults were performed, as were the six analyses for control during the reservoir leak fault. The significant contrast is listed on the right side of Table 3.

Boundary 1: Low-Level Data

Of the 3200 probes administered, 136 were discarded (4.25%). The significant preplanned contrasts for both accuracy and latency are listed on the left side of Table 3; the means for each display are illustrated in Figure 9b.

DISCUSSION

In Experiment 1, we found that the digital display produced the poorest performance for Boundary 3 tasks. Similar results were obtained in Experiment 2. Two contrasts indicated that the digital display produced significantly degraded control performance during nonfault trials (see Figure 9a). Contrast 10 (standard deviation of the error) indicated that control performance with the digital display was significantly more variable than was control performance with all other displays. Contrast 9 (acquisition time) indicated that it took significantly longer for the indicated level to reach the target band with the digital display than with the composite display.

In addition to this finding, there were other indications that the composite display improved control and fault detection performance significantly (also consistent with the findings obtained in Experiment 1). Contrast 11 revealed that the indicated level initially reached the target band in less time with the composite display than with all other displays. Contrast 11 also indicated that the false alarm rate was significantly lower with the composite display than with all other displays.

In Experiment 1, the Boundary 1 evaluation revealed that the presence of scale markers/gridlines and digital values improved performance significantly for low-level data probes with little or no tradeoffs. Also the presence of color-coding/layering/separation was not found to influence performance significantly.

A similar pattern was obtained in Experiment 2 for these three design techniques. Digital values (digital and composite displays) improved the participants' ability to perform low-level data probes, as indicated by the significance of Contrasts 8, 10, and 11 for both accuracy and latency (see Figure 9b). Contrast 9 for

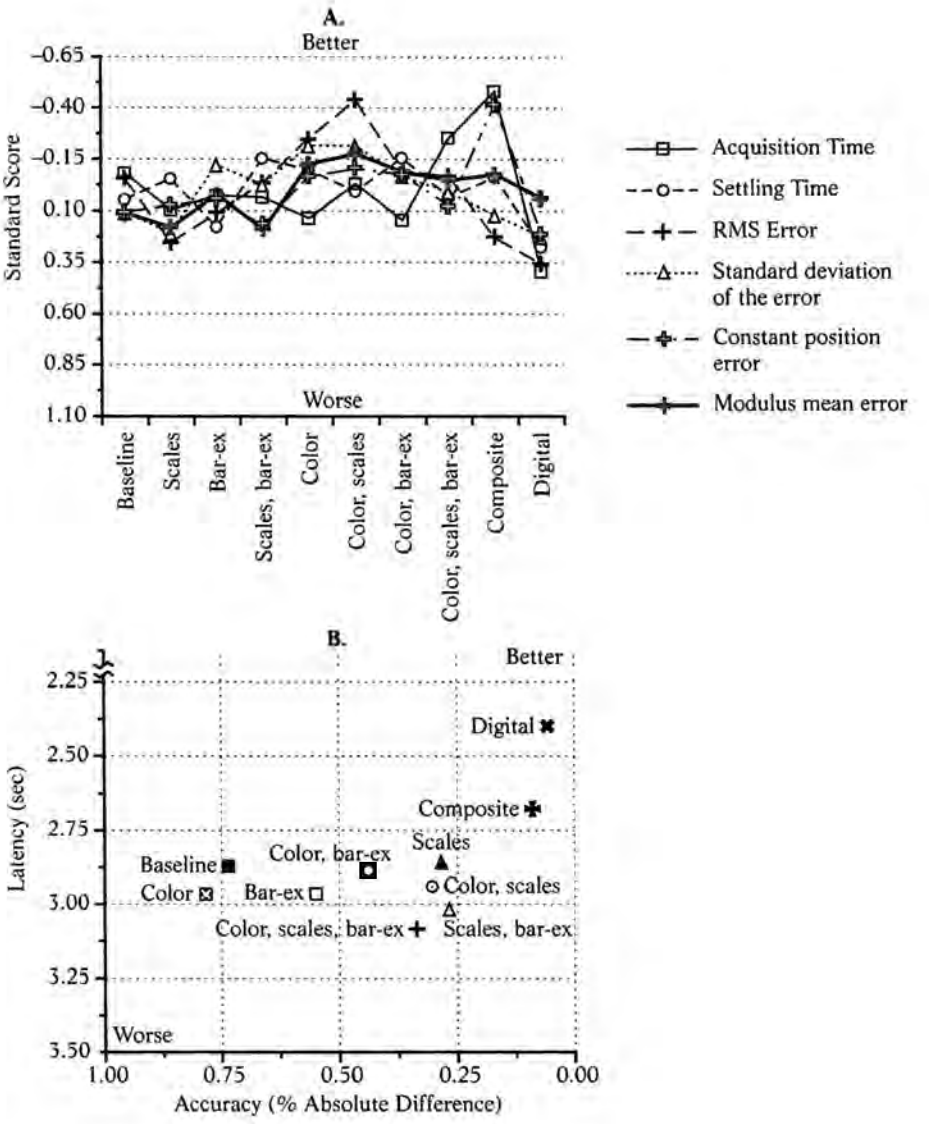

Figure 9. Results obtained for displays in Experiment 2. (a) Boundary 3 evaluation: average performance (standard scores) for the six measures of system control during nonfault trials. (b) Boundary 1 evaluation: latency (in seconds) and accuracy (error magnitude) for the low-level data probes.

latency again revealed that the digital display was more effective in supporting performance for low-level data probes than was the composite display. Contrasts 1, 4, and 5 all indicated that the presence of scale markers/gridlines improved the accuracy of probe estimates significantly. The color design technique was involved in two significant interaction contrasts (Contrast 6 for accuracy and latency), but no significant benefits or costs were apparent.

In contrast to these three techniques, a different pattern of results was obtained for the bar-ex design technique in Experiment 2. In Experiment 1 the application of this technique degraded performance significantly; in Experiment 2 the application of this technique improved performance significantly (Contrasts 3, 5, and 6 for accuracy). The bar-ex design tnique did tend to increase response latencies (Contrast 6 for latency), but the differences were not significant.

A comparison of results across Boundaries 1 and 3 yields overall conclusions similar to those outlined for Experiment 1 (see Table 3). If a display manipulation produced significant differences at both boundaries, the patterns of

performance were usually in the opposite direction (i.e., the contrasts labeled with "B" superscripts). As in Experiment 1, there was evidence (Contrasts 9 and 10) that the digital display improved performance at Boundary 1 (accuracy and latency) but degraded performance at Boundary 3 (acquisition time and standard deviation of the error). Only one contrast revealed a similar pattern of results across boundaries (Contrast 11). As in Experiment 1, the composite display improved performance significantly at both boundaries (as indicated by the contrasts labeled with "C" superscripts).

GENERAL DISCUSSION

Interpretation of Boundary 1 Results

The results of Experiments 1 and 2 indicate that three of the four design techniques applied to the configural display improved performance at the Boundary 1 evaluation (i.e., the low-level data probes). These results will be interpreted in terms of the mutually interacting constraints described in the introduction (task, display, observer). The task constraints were to provide a quantitative estimate of an individual variable. The primary observer constraints were related to basic cognition/perception/action capabilities and were therefore reasonably similar across participants.

The results indicate that the constraints imposed by the various display conditions were different; those constraints associated with the baseline configural display (Figure 1) will be considered first. The visual features relevant to the probe task were the appropriate data marker (i.e., a side of the rectangle) and the numeric labels denoting scale on the appropriate axis. The fact that these visual features were separated in space introduced a difficult set of constraints, forcing participants to complete two mental estimates and a mental computation to produce a response.

The first mental estimate involved an extrapolation from the spatial position of the data marker to an inferred spatial position on the appropriate axis. This location then had to be considered with regard to the two closest axis labels (representing a 10% increment in the axis scale). The second mental estimate involved the derivation of a value corresponding to the portion of the 10% increment between the inferred location of the data marker and an axis label. This value then had to be added to (or subtracted from) the value of an axis label (a mental computation) to derive the final estimate.

Three of the four design techniques added visual information to the baseline condition, changing the nature of the imposed display constraints. This visual information provided a better match to observer constraints by allowing powerful perceptual processes to replace mental processes. The bar-graph/extender (bar-ex) technique, as it was implemented in Experiment 2, improved performance because it eliminated the first mental estimate: The extender lines were superimposed on the axis scale. Thus there was no need to mentally extrapolate from data marker to axis. The scales design technique improved performance because the gridlines projected the axis scale into the display area, also eliminating the first mental estimate. In addition, the gridlines provided a more precise indication of scale than did the numeric labels, making the second mental estimation (the portion of the 10% increment between scale markers) an easier task. Providing digital values matched the constraints of the task exactly; thus the value of the individual variable was available directly, and there was no need for any mental estimation.

The color technique did not improve performance because the visual structure it provided (chromatic and luminance contrast) did not eliminate either mental estimate. Please note that alternative input devices could change the nature of the interface constraints in a similar fashion. For example, entering estimates with an analog slider (rather than typing numbers) could provide additional visual structure useful in eliminating mental estimates (i.e., aligning visual components).

Interpretation of Boundary 3 Results

In the Boundary 3 evaluation, the primary task constraints are those associated with the manual control of feedwater simulation. The participant must not only be able to obtain information from the display but must also know how to utilize that information to perform system control and fault detection tasks.

The results of the Boundary 3 evaluation revealed considerable differences in the ability of the various displays to support participants in the completion of these tasks. A primary finding was that performance with the digital display (Figure 6) was particularly poor. This is clearly illustrated in Figures 7a and 9a and in the statistical analyses. All but one of the significant contrasts (Contrast 1, Experiment 1; see Table 2) for the Boundary 3 evaluations included this display, and in each case it was associated with degraded performance.

Therefore, the interpretation of these results will rest on an analysis and comparison of the differences in display constraints that were imposed by the digital display and the basic analog configural display (the geometrical form present in all other displays). The analog configural format supports performance at this boundary because it represents the critical domain semantics directly. For example, the height and width of the rectangle (in combination with the coding conventions of the individual variables) are highly salient emergent features that correspond to the critical system properties of energy and mass balance, respectively. The shape, size, and location (position in the display grid) of the rectangle provide additional emergent features that testify to the current system state.

In addition, these critical properties can be viewed in the context of system goals (e.g., how close is the indicated level to the goal or trip set points?). In essence, participants could utilize powerful pattern recognition capabilities to assess the current system and to help determine the correct control input.

In contrast, the digital display imposed a severe set of constraints. The route to underlying meaning was much less direct: The domain semantics (relationships, properties, goals, and constraints) were not directly visible. Instead, the participants were forced to derive this information mentally using the digital values in conjunction with their knowledge about the system. (See Bennett & Flach, 1992, and Bennett et al., 1997, for a more detailed discussion of similar considerations.) Thus the constraints introduced by the digital display made it much more difficult to assess the system state, determine the appropriate control input, and gauge the appropriateness of the system dynamics. As a result, performance suffered.

These results should not be interpreted as evidence that the analog configural display is an optimal one. For example, Bennett et al. (1997) presented an alternative display suitable for use in this domain (using a design logic similar to Vicente, 1991) that presents information from all five levels of Rasmussen et al.'s (1994) abstraction hierarchy. In contrast, the present configural display provides very little information about physical processes or physical form. This is a clear limitation. Information regarding goals and abstract function is represented directly (e.g., mass and energy balance as the width and height of the rectangle, respectively). However, the coding conventions associated with individual variables must be considered in order to relate these emergent features to the system state unambiguously. The potential limitations of this aspect of the display are currently being explored.

Generalization between Evaluation Boundaries

Rasmussen et al.'s (1994) evaluative framework suggests that the generalization of results between two boundaries will occur for a display only when it produces a set of visual constraints matching the task and observer constraints that exist at both boundaries. The results of the present experiments appear to be very consistent with these suggestions. A diametrically opposed pattern of results was obtained for the digital display at the two boundaries. The Boundary 1 results revealed that the digital display produced the best performance at low-level data probes; the Boundary 3 results revealed that the digital display produced the worst system control performance (see contrasts with "B" superscripts in Tables 2 and 3). These results strongly suggest that a fundamentally different set of task and observer constraints existed at the Boundary 1 and Boundary 3 evaluations. Furthermore, they suggest that the display constraints imposed by the digital display were well matched to the task/observer constraint envelope at Boundary 1 but not to the task/observer constraint envelope at Boundary 3.

In contrast, the composite display was effective in supporting performance at both boundaries

(see contrasts with "C" superscripts in Tables 2 and 3). This display produced two of the three instances in which a contrast revealed the same pattern of results across boundaries (see Contrast 11 in Tables 2 and 3). It is unlikely that the generalization of results across boundaries resulted from a single design feature or a higher-level property arising from the combination of design features; otherwise, it would have been revealed directly in other contrasts. The most likely interpretation is that participants could select and use the specific design features in the composite display that were appropriate for tasks at each boundary. This interpretation is based on the independent results at each boundary level: The analog configural display supported performance at Boundary 3; the scales, bar-ex, and digital value design techniques supported performance at Boundary 1.

This is an encouraging set of results for configural display design. As outlined in the introduction, one potential disadvantage associated with the use of configural displays involves the ability of participants to obtain low-level data. The results indicate that design features can be combined in a single display to support performance at a number of boundary levels with relatively little interference. The results represent progress toward a fundamental display design goal: single graphical displays capable of supporting performance at multiple tasks. Hansen (1995, p. 542) foreshadowed these results in stating that "human factors researchers should not treat the discussion of graphical versus analytical (e.g., numerical) interfaces as an either/or issue. Instead, they should be studying ways to improve the integration of these interfaces."

One final observation is that the collective results provide a fairly clear message with regard to the generalization of results between boundaries. Despite the hundreds of contrasts performed during the statistical analyses, there was very little evidence to support generalization. The design features that improved performance significantly at Boundary 1 had very little positive impact on the performance of more complex domain tasks at Boundary 3. In fact, there was only one pair of significant contrasts that provided unequivocal evidence supporting the generalization of results across boundaries: Contrast 1 in Experiment 1 (see Table 2). This pattern of results was not replicated in Experiment 2.

Additional Issues with Evaluation Boundaries

Global considerations regarding evaluations at alternative boundary levels will now be considered. Rasmussen et al. (1994) emphasized that fundamentally different methodological approaches may be required at alternative levels of evaluation. They state, "it is clear that the empirical approach to evaluations has its limitations. It seems best suited for separate tasks or functions [lower boundaries]. ...For more complicated situations [higher boundaries] ... the empirical approach is most difficult to carry out convincingly and realistically because of all the uncontrolled (uncontrollable) variables" (pp. 209–210). For evaluation at higher-level boundaries, a more qualitative or process-oriented approach may be required: "Detailed analyses of the individual trajectories and generalizations across samples are more important than quantitative data" (p. 209). Our findings are consistent with these observations.

The results of the two experiments indicate that the Boundary 1 tasks were much more amenable to controlled laboratory experimentation than were the Boundary 3 tasks. For example, only 3% of the contrasts performed at Boundary 3 were found to be statistically significant, whereas 52% of the Boundary 1 contrasts were significant. It is possible that a stand-alone Boundary 1 evaluation (i.e., one that was not embedded within a Boundary 3 evaluation) would have made these differences more pronounced because of increased experimental control.

Several factors might have contributed to the lack of significant results at Boundary 3. One simple possibility is that the smaller number of participants reduced the statistical power. A second possibility is that alternative displays could have improved performance at Boundary 3. A third possibility is that the dependent measures used for the Boundary 3 control and fault detection tasks were not particularly sensitive. Alternative methodologies (e.g., verbal protocols, state space, state transition diagrams, etc.; see Moray, Lootsteen, & Pajak,

1986; Sanderson, Verhag, & Fuld, 1989; and Yu, Lau, Vicente, & Carter, 1998) may have provided more sensitivity.

Although there are alternative explanations, we believe that the overall pattern is largely a result of fundamental differences that exist between the evaluation boundaries. The degree of experimental control that can be imposed in a Boundary 3 evaluation is less than that for a Boundary 1 evaluation. As a result, the likelihood of statistically significant differences is reduced. A number of factors contribute to a lower level of experimental control. The Boundary 3 tasks were inherently more difficult to complete and presented greater challenges than the Boundary 1 tasks. The role of basic perception/cognition/action capabilities (relatively similar across individuals) played a primary role in the completion of Boundary 1 tasks. In contrast, task-specific knowledge and skills (more likely to vary across individuals) played a primary role in the completion of Boundary 3 tasks. Many more degrees of freedom in the action alternatives existed at Boundary 3; the potential for the development of alternative strategies (either more or less effective) was therefore greater.

This potential was fully realized in the Boundary 3 evaluation. All participants in the two experiments were able to perform the Boundary 1 tasks effectively. However, only six of the nine participants were able to perform the Boundary 3 tasks effectively. Figure 10a summarizes the overall differences in control performance between effective and ineffective participants during nonfault trials of Experiment 1. The two lines represent a linear best fit for performance across experimental sessions. Note that the effective participants (open symbols, gray line) improved their performance as they became more familiar with the task, whereas the ineffective participants (filled symbols, black line) did not.

Detailed analysis of each participant's control performance (one example of the qualitative approach referred to by Rasmussen et al., 1994) indicated why these large differences in performance occurred. Effective controllers not only understood system dynamics (e.g., counterintuitive shrink and swell effects resulting from thermodynamic properties) but also used these dynamics in the pursuit of system goals. In particular, all five effective controllers made extensive use of thermodynamic control "levers."

Figure 10b presents average values of feed flow (the variable controlled directly) as a function of time into trial during all nonfault trials. The thin lines represent each participant's performance; the thick lines represent averages for effective and ineffective participants. Effective controllers (thick gray line) decreased feed flow as quickly as possible from approximately 30 to 70 s into the trials (see arrows in Figure 10b). This produced thermodynamic swell effects that moved the indicated level toward the goal value quickly and effectively. None of the three ineffective controllers (black lines) developed this strategy. Similar observations have been useful in differentiating between levels of expertise in real-world feedwater operators. For example, Roth, Woods, and Gallagher (1986, p. 180) stated that expert operators "formulate response strategies that exploit the shrink and swell characteristics of the process to their advantage. Particularly characteristic is the use of shrink and swell as a source of control."

Summary

Rasmussen et al.'s (1994) framework emphasizes that interface and display design requires iterative evaluation at multiple levels. The evaluations conducted at each boundary level will have unique goals and will provide unique insights for design. The Boundary 1 evaluations in the present studies focus on issues in physical form. More precisely, the focus is on the relationship between basic human cognition/perception/action capabilities and specific graphical features. The primary concern is whether an observer can effectively obtain information from a display. A Boundary 1 evaluation of this type need not concern only low-level data (as in the present studies). For example, this type of Boundary 1 evaluation might also be particularly useful in evaluating alternative configural display designs that use different visual features for the representation of critical domain properties. Evaluations at this level are important because they provide the groundwork for building effective displays.

However, effective design also requires

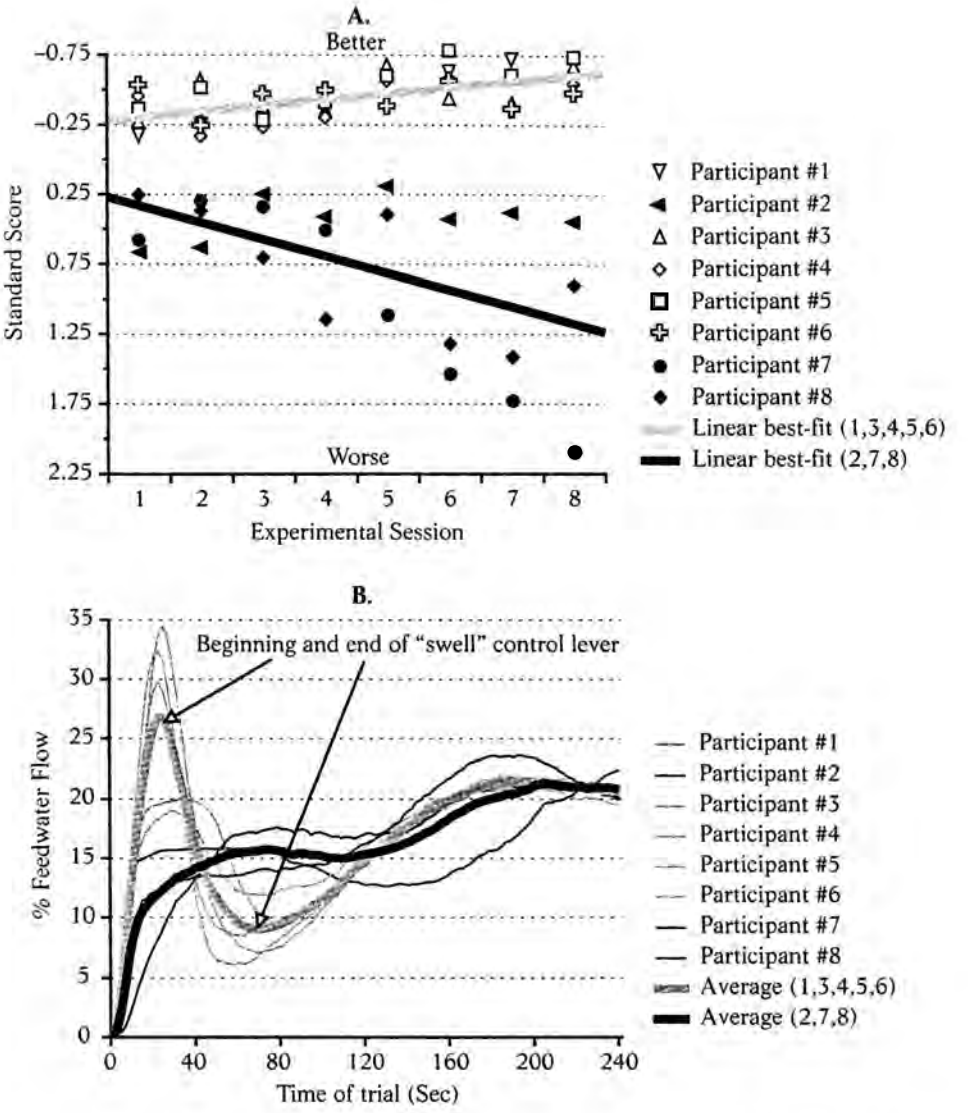

Figure 10. Control performance for all participants in the nonfault trials of Experiment 1. (a) Each symbol represents the average standard score (across the six measures of control performance) for each participant and experimental session; scores plotted higher in the graph represent better performance. Each line represents a linear best fit for groups of participants who were effective (open symbols and gray line) or ineffective (filled symbols and black line) at control tasks. (b) The average values of feedwater flow as a function of time into trial for all nonfault trials. Each thin line represents the average performance for an individual participant; the thick lines represent average performance for effective (gray lines) and ineffective (black lines) controllers.

going beyond the relationship between graphical form and perception. At higher levels of evaluation, the domain constraints will play an important role. The information that can be obtained easily must also be semantically meaningful in the context of the domain task(s) to be performed. Thus Boundary 3 evaluations work in concert by assessing the extent to which display constraints are a reflection of domain constraints. Only when this occurs will the display support the completion of more complex domain tasks.

The lack of generalization between boundary levels in the present experiments, the tenets of Rasmussen et al.'s (1994) framework, and various commentaries (e.g., Vicente, 1997) all reinforce the observation that there is a need to test the effectiveness of interface and display solutions at levels of evaluation that more closely mirror the complexities encountered in

applied settings. Striking a balance between the desire for scientific rigor and the needs of an applied discipline is truly a formidable challenge for display designers and the human factors community in general.

ACKNOWLEDGMENTS

The authors thank David Woods, John Flach, William Howell, Richard Jagacinski, and two anonymous reviewers for discussions and comments on earlier drafts. Funding was provided by Wright State University and the Ohio Board of Regents (Research Challenge and Research Incentive Grants). The part-task simulation was originally developed at Westinghouse.

REFERENCES

Bennett, K. B., & Flach, J. M. (1992). Graphical displays: Implications for divided attention, focused attention, and problem solving. *Human Factors, 34,* 513–533.

Bennett, K. B., Nagy, A. L., & Flach, J. M. (1997). Visual displays. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 659–696). New York: Wiley.

Bennett, K. B., Payne, M., Calcaterra, J., & Nittoli, B. (2000). An empirical comparison of alternative methodologies for the evaluation of configural displays. *Human Factors, 42,* 287–298.

Bennett, K. B., Toms, M. L., & Woods, D. D. (1993). Emergent features and configural elements: Designing more effective configural displays. *Human Factors, 35,* 71–97.

Hansen, J. P. (1995). An experimental investigation of configural, digital, and temporal information on process displays. *Human Factors, 37,* 539–552.

Keppel, G. (1982). *Design and analysis. A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lovie, P. (1986). Identifying outliers. In A. D. Lovie (Ed.), *New developments in statistics for psychology and the social sciences* (pp. 44–69). London: British Psychological Society and Methuen.

Moray, N., Lootsteen, P., & Pajak, J. (1986). Acquisition of process control skills. *IEEE Transactions on Systems, Man and Cybernetics, SMC-16,* 497–504.

Poulton, E. C. (1974). *Tracking skill and manual control.* New York: Academic.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering.* New York: Wiley.

Roth, E. M., Woods, D. D., & Gallagher, J. M., Jr. (1986). Analysis of expertise in a dynamic control task. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 179–182). Santa Monica, CA: Human Factors and Ergonomics Society.

Sanderson, P., Verhag, A. G., & Fuld, R. B. (1989). State-space and verbal protocol methods for studying the human operator in process control. *Ergonomics, 32,* 1343–1372.

Vicente, K. J. (1991). *Supporting knowledge based behavior through ecological interface design* (Tech. Report EPRL-91-1). Urbana-Champaign: University of Illinois, Engineering Psychology Research Laboratory and Aviation Research Laboratory.

Vicente, K. J. (1997). Heeding the legacy of Meister, Brunswik, and Gibson: Toward a broader view of human factors research. *Human Factors, 39,* 323–328.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37,* 473–494.

Yu, X., Lau, E., Vicente, K. J., & Carter, M. W. (1998). Advancing performance measurement in cognitive engineering: The abstraction hierarchy as a framework for dynamical systems analysis. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 359–363). Santa Monica, CA: Human Factors and Ergonomics Society.

Kevin Bennett is an associate professor in the Department of Psychology at Wright State University in Dayton, Ohio. He received a Ph.D. in Applied-Experimental Psychology from the Catholic University of America in 1984.

Brett Walters is a human factors engineer at Micro Analysis & Design, Inc., in Boulder, Colorado. He recieved an M.S. in Human Factors Psychology from Wright State University in 1997.

Date received: April 20, 1998
Date accepted: September 22, 2000