An Empirical Comparison of Alternative Methodologies for the Evaluation of Configural Displays

Kevin B. Bennett, Michael Payne, Jeffery Calcaterra, and **Bob Nittoli,** Wright State University, Dayton, Ohio

Two different methodologies (visual, memory) were used to evaluate alternative versions of the same configural display. One version (composite display) had several graphical design techniques applied, whereas the other version (baseline display) did not. Two types of information probes (high-level property, low-level data) were administered. When the displays were visible during completion of the probes (visual methodology), the display manipulation had the largest impact on performance (composite display associated with better performance); when the displays were not visible (memory methodology) the probe manipulation had the largest impact on performance (high-level probes associated with better performance). These results are interpreted in light of the mutually interacting constraints introduced by factors in display design, task requirements, and the participants' cognitive and perceptual capabilities/limitations. Implications for both the design and the evaluation of displays and interfaces in general are discussed. Actual or potential applications of this research include design techniques for improving the quality of graphic displays and methodological insights for interpreting previous research and guiding future experimentation.

INTRODUCTION

For several years, an ongoing research project has focused on issues in the design and evaluation of graphical displays. This research has provided both empirical (Bennett, Toms, & Woods, 1993; Bennett & Walters, in press) and theoretical (Bennett & Flach, 1992; Bennett, Nagy, & Flach, 1997) contributions to the literature. The empirical studies have been focused on issues in the design of *configural displays*, a display format that involves the mapping of several individual variables into a single geometrical form (the form changes shape dynamically as a function of changes in the individual variables). The polar graphic format of Woods, Wise, and Hanes (1981), which maps numerous variables into an octagonal-shaped geometric form, is a well-known example. These studies have also explored issues in evaluation in that the impact of the various display manipulations on performance were assessed concurrently with multiple methodologies. The focus of the current paper is on methodological issues; we provide a brief description of the issues in design to set the stage.

A great deal of laboratory research has investigated issues in the design of configural displays. One primary concern has been with their performance trade-offs relative to other display formats, especially *separable displays* (in which each variable has a unique graphical representation such as a bar graph). Benefits and costs are often assessed using tasks that are located at various points along a continuum. The tasks range from those that require the consideration of individual variables to those that require consideration of relationships among variables. The end points of this continuum can be referred to as *low-level data* and *high-level property*; the former refers to "local constraints or elemental state variables

Address correspondence to Kevin B. Bennett, Psychology Department, 335 Fawcett Hall, Wright State University, Dayton, OH 45435; kevin.bennett@wright.edu. **HUMAN FACTORS**, Vol. 42, No. 2, Summer 2000, pp. 287–298.

that might be measured by a specific sensor," and the latter refers to "more global constraints that reflect relations or interactions among multiple variables" (Bennett et al., 1997, p. 683).

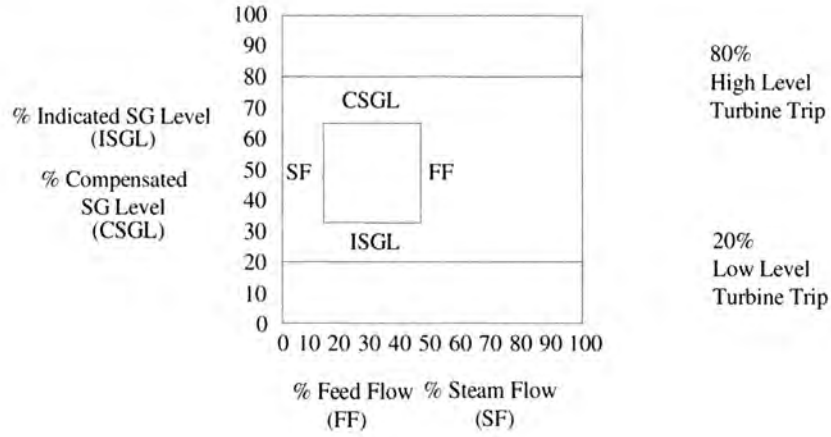
Two recent articles (Bennett & Flach, 1992; Wickens & Carswell, 1995) have reviewed this literature and have drawn fairly similar conclusions with respect to the pattern of results that has emerged. In general, configural displays are more effective than separable displays for the performance of high-level property tasks. Configural displays capitalize on powerful pattern-recognition capabilities and allow observers to assess the state of the underlying domain directly by perceiving the patterns of distortion relative to the prototypical geometric form. The pattern is less clear for low-level data tasks. Although the most common finding is a lack of statistical significance, when significant differences are found between display types, they tend to favor separable formats. The results of Bennett et al. (1993) are representative: (a) Participants responded more quickly and significantly more accurately to high-level property probes with a configural display than with a separable display, and (b) the relative effectiveness of the two types of displays was reversed for low-level data probes (significant under some conditions and nonsignificant under others).

A single graphical format that is capable of supporting performance at both types of tasks is preferable to multiple formats (each specialized for a particular type of task) for a variety of reasons. The amount of display "real estate" is often limited, and multiple formats require additional navigational/selection input from the observer. We pursue this design goal in the present experiment by evaluating several design strategies aimed at improving the accessibility of information encoded into configural displays. Two versions of the same basic configural display were developed for a simulated control task (the manual control of feedwater). In the *baseline* configural display, the value of four system variables was mapped into a rectangle that could change shape, size, and location in the display grid (Figure 1a). In the *composite* display (Figure 1b), several design strategies were applied to the baseline display, including scale markers/grid lines, color-coding/layering/separation, digital values, and graphical extenders (see Bennett et al., 1993, for a more complete description of the design rationale for the display).

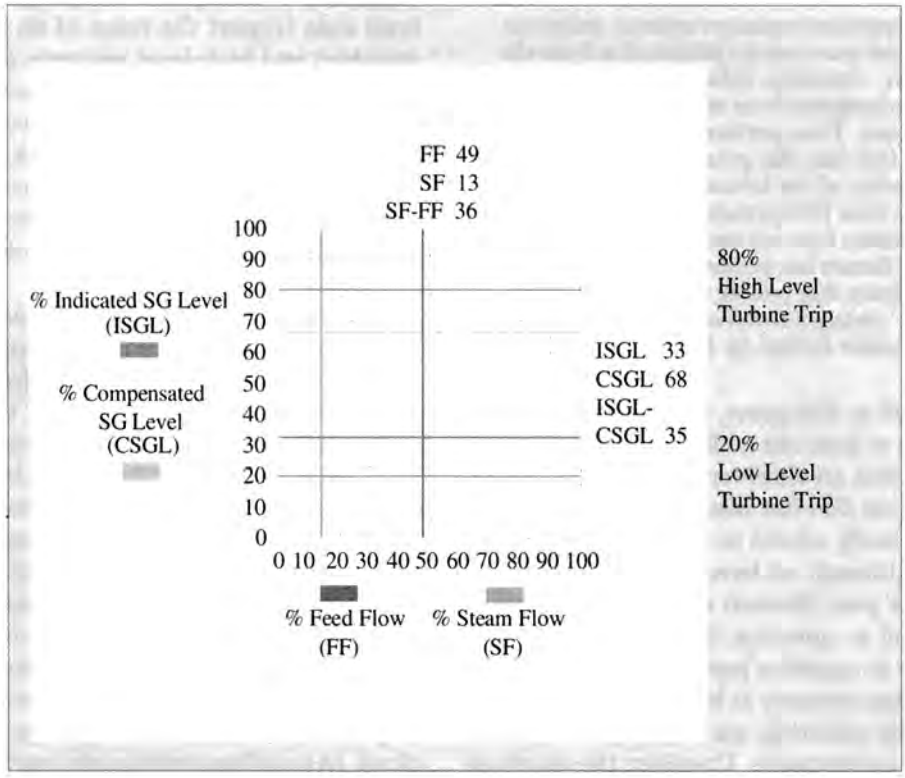
In addition to completing the control task, participants completed both low-level data and high-level property information probes. For the low-level data probes, participants had to provide estimates of one of the four variables. For the high-level property probes, participants had to provide estimates of the absolute differences between two sets of variables (steam flow vs. feed flow or indicated steam generator level vs. compensated steam generator level). These differences corresponded to the system properties of mass balance and energy balance, respectively.

The methodological manipulation of primary interest involved the use of two alternative methodologies to evaluate performance. With the visual methodology, the displays were visible during the completion of low-level data and high-level property probes. This allowed observers to use perceptual systems to extract information from the two displays. Studies employing similar visual methodologies are most common in the literature on static graphs (Cleveland, 1985; Cleveland & McGill, 1985; Gillan & Lewis, 1994; Gillan & Richman, 1994; Meyer, Shinar, & Leiser, 1997). With the memory methodology, the displays were removed from sight, thus requiring an observer to use memory systems to complete the two probes. Barnett and Wickens (1988), Bennett et al. (1993), and Wickens and Andre (1990) employed similar memory methodologies.

These two categories of methodologies can be used in very similar ways, as illustrated by the direct correspondence between the high-level property and low-level data memory probes employed by Bennett et al. (1993) and the identification and subtraction visual tasks employed by Gillan and Lewis (1994). However, there are fundamental differences in the underlying rationale and assumptions associated with the methodologies. The rationale behind the visual methodology is fairly straightforward. A display will facilitate information extraction performance to the extent that the information has been encoded in a manner



A



B

Figure 1. The two configural displays used in the experiment. In both displays four variables are mapped into a configural display with the geometric form of a rectangle. In the baseline display (a) all graphical and textual elements are presented at roughly the same levels of visual salience or prominence. The composite display (b) had four design techniques applied to provide increased visual structure and information: scales, color coding/layering/separation, extenders, and digital values.

that is consistent with the perceptual capabilities and limitations of the observer. Therefore, better information extraction performance constitutes evidence that one fundamental concern in effective display design has been met. The assumptions for this methodology are few. For example, researchers might assume that the ability to extract information from a display will generalize to the ability to perform more complicated domain tasks. This need not be true, because the information that is easily extracted must also be meaningful in terms of the domain semantics that underlie these tasks.

The assumptions and rationale of the memory methodology are fundamentally different. The following quote by Bennett et al. (1993) provides an example of this difference:

> To complete retrospective memory probes an individual must extract information from the displays, represent information internally, recall information from memory, and generate a response. Thus, performance on a memory probe task has the potential to reveal the availability of the information in a graphic display. One interpretation of differences in performance between two displays is that one display format has presented the information in a manner that is more compatible with observers' perceptual and cognitive capabilities than another format. (p. 86)

As implied in this quote, the memory methodology has at least one additional assumption: It assumes that an observer's capability to recall information that has been presented in a display is directly related to the quality of display design. Although we have used this methodology in the past (Bennett et al., 1993), we feel compelled to question this assumption. Researchers in cognitive psychology have demonstrated that memory is both far from perfect and highly selective, particularly for certain types of information. Consider the *levels of processing* theory of memory (Craik & Lockhart, 1972). This theory maintains that stimuli can be processed at different levels and that the strength of memory is determined by the depth of that processing. For example, information about words can be encoded on the basis of their physical appearance, what they sound like, or what they mean. Generally speaking, the quality of recall has been shown to be much better for the gist or meaningful aspects of a stimulus than for its physical details (e.g., Bartlett, 1932; Bransford & Franks, 1972; Parkin, 1984). The critical implication for display evaluation is that when participants are required to remember information about the physical details of a display, the results that are obtained may be attributable to limitations in memory and to the quality of the display design.

In the present study we investigate these and related concerns. Participants performed a simulated control task (the manual control of feedwater) and were interrupted occasionally so that we could assess their capability to either extract (visual methodology) or remember (memory methodology) information presented in the baseline or composite display (Figure 1). Participants completed both low-level data (report the value of an individual variable) and high-level property (report the absolute value of the difference between two variables) probes. Measures of accuracy and latency were recorded. Thus the experimental design included four within-subjects factors: display (baseline, composite), probe (low-level data, high-level property), methodology (memory, visual), and session (1–8).

We predicted that performance at the probe tasks would be better for the composite display than for the baseline display: The design strategies provided both exact digital values and additional structure (a visual context) that were likely to improve the accessibility of the information that was to be reported. We also predicted that performance advantages for the composite display would be more pronounced with the visual methodology than with the memory methodology. Although the physical details required to complete probes could be accessed directly in the visual methodology, they had to be recalled in the memory methodology. (As outlined previously, memory is not particularly good for this type of information.) Similarly, probe type was expected to have a more pronounced impact on performance with the memory methodology because the two probe types were mapped into very different visual features. High-level properties were represented directly as the height or width of the rectangular form; low-level data were represented indirectly as the distance between a side

of the rectangle and a display axis. We expected that the former visual features would be more memorable than the latter features and that performance at the high-level property probes would therefore be facilitated.

METHOD

Participants

The participants consisted of eight students (four men and four women; five graduate and three undergraduate psychology majors) who had normal or normal-corrected vision and normal color perception. Participants were paid $5.00/h for their participation.

Apparatus

A Sun Microsystems 4-110 Workstation was used, which includes a color video monitor (40.64 cm, 1152 × 900 pixel resolution), a standard keyboard, and a three-button optical mouse.

Simulation Model

The part-task simulation replicated the basic dynamic characteristics of a single nuclear power plant steam generator during start-up. The manual control of feedwater task involved the control of mass flowing into (feedwater flow, or FF) and out of (steam flow, or SF) a steam generator so that the level of coolant inside (indicated steam generator level, or ISGL) was maintained between upper and lower limits. Control was complicated by the fact that energy inflow/outflow also influenced ISGL at a different time constant (i.e., by producing counterintuitive shrink/swell effects). Compensated steam generator level (CSGL) was a calculated predictive variable (a form of decision support) that provided an estimate of ISGL that was not confounded by these thermodynamic effects (see Bennett et al., 1993, for a more detailed description of the simulation model).

Displays

Two displays, baseline and composite, were used (see Figure 1). Both displays mapped the four variables into the geometrical form of a rectangle (Figure 1). ISGL and CSGL were plotted on the vertical axis; SF and FF were plotted on the horizontal axis. Each variable had a label that was positioned close to the appropriate side of the rectangle and moved with the rectangle. Each axis was 10.16 cm and subtended a visual angle of 11.60 (assuming that an observer sat 50 cm from the screen). The displays were updated every 2 s. In the baseline display, all graphical elements were represented at approximately the same level of salience (Figure 1a). In the composite display, four display design techniques were applied (Figure 1b). The color/layering/separation technique included the use of gray scale shading for the three display mattes and color coding (blue, yellow, green, and purple) for the sides of the rectangle. The extender technique consisted of linear extensions from the sides of the geometric form (i.e., the rectangle) to the appropriate axes (which emphasized the contribution of individual variables relative to the scale). The scales technique provided grids that crossed the display area horizontally and vertically at 10% intervals. The digital display technique involved annotating the analog configural display with digital values for each variable, energy balance (ISGL vs. CSGL), and mass balance (SF vs. FF).

Procedure

Participants were tested individually during a 2-week period and completed nine 1-h sessions. The first session was used for training; participants received both written and oral instructions describing the simulation and the tasks. The experimenter remained in the room and answered general questions. In each of the ensuing eight experimental sessions, the participants completed two blocks of trials (one for each display). The order of display presentation was counterbalanced between participants and sessions. Within each block there were approximately six experimental trials; each lasted up to 5 min. During a trial participants completed both the manual control of feedwater task and information probes.

Virtually all of the details of the control task were as described in Bennett et al. (1993). Participants increased or decreased the rate of FF and were instructed to maintain ISGL between the upper (20%) and lower (80%) set points for as long as possible. Time on task

was measured from the beginning of a trial until ISGL crossed one of the set points or until the end of a 5-min trial. The four starting positions for ISGL were 35%, 45%, 55%, and 65%. In addition, control was made progressively more difficult through the course of a trial by the introduction of both continuous and asynchronous changes to the steam flow parameter. The continuous changes consisted of three steam flow ramp types: oscillating (null ramp), oscillating with a gradual rise (rising ramp), or oscillating with a gradual fall (falling ramp). The asynchronous changes consisted of random disturbances to steam flow. The severity of the continuous and asynchronous changes varied as a function of the elapsed time of an experimental trial.

The probes occurred at random times during a trial and were administered when the next screen update was scheduled to appear. An auditory tone sounded, and a description of the probe was presented in a text panel at the top of the screen (e.g., "Enter % value for SF"). The participants entered a numeric response on the keyboard; they were instructed to respond as accurately and as quickly as possible. Accuracy scores were computed by taking the absolute value of the difference between the observer's estimate and the actual value that appeared on the screen. Latency was measured from the time that the prompt appeared until the participant entered the first digit of their response (0.01 s accuracy). Participants were given feedback on accuracy.

Two types of probes were administered: low-level data and high-level property. A low-level data probe required an observer to enter the value for one of the four individual variables; a high-level property probe required participants to enter the absolute value of the difference between either (a) ISGL and CSGL (energy balance) or (b) SF and FF (mass balance). Two types of methodologies were used to administer probes: visual and memory. In the memory methodology, the screen was blanked before the probe was presented; this forced the observer to recall the information from memory. In the visual methodology, the screen was not blanked before the probe (i.e., the display remained available during a probe). An algorithm ensured an approximately equal distribution of probes among the various conditions.

RESULTS

Tests for outliers (accuracy, latency, and time on task) were performed using standardized deviate statistics (Barnett & Lewis, 1984; Lovie, 1986; Ratcliff, 1993). The test is described in Lovie (1986, pp. 55–56): $T_1 = (x_{(n)} - \bar{x})/s$, where $x_{(n)}$ 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. An outlier status for either accuracy or latency resulted in the removal of both scores. Of the 5503 probes that were administered, 131 were discarded (2.38%); Wilcoxon signed ranks tests revealed that the distribution was random. Of the 726 time-on-task scores, 3 were discarded (0.41%). Analyses with and without outliers revealed the same general pattern of significant findings; the minor differences do not change the nature of the conclusions that have been drawn.

Accuracy

Scores were averaged across experimental session, and a 2 (display) × 2 (probe) × 2 (method) repeated-measures analysis of variance (ANOVA) was performed. The main effects of display, $F(1, 7) = 48.16$, $p < .0003$, method, $F(1, 7) = 186.43$, $p < .000003$, and probe, $F(1, 7) = 30.47$, $p < .0009$, and the Display × Method, $F(1, 7) = 11.71$, $p < .02$, Display × Probe, $F(1, 7) = 6.27$, $p < .05$, Method × Probe, $F(1, 7) = 56.04$, $p < .0002$, and Display × Method × Probe, $F(1, 7) = 12.00$, $p < .02$, interactions were significant. The means for the three-way interaction are illustrated in Figure 2. Planned comparisons to assess performance differences between the composite and baseline displays were conducted for each methodology/probe combination. In Figure 2 these comparisons are identified by pairs of filled versus open symbols that have a similar shape and methodology (e.g., the symbols labeled 1 and 2). For the visual methodology, the composite display improved performance significantly for both low-level data (1 vs. 2), $F(1, 7) = 106.58$, $p < .00002$, and high-level

was measured from the beginning of a trial until ISGL crossed one of the set points or until the end of a 5-min trial. The four starting positions for ISGL were 35%, 45%, 55%, and 65%. In addition, control was made progressively more difficult through the course of a trial by the introduction of both continuous and asynchronous changes to the steam flow parameter. The continuous changes consisted of three steam flow ramp types: oscillating (null ramp), oscillating with a gradual rise (rising ramp), or oscillating with a gradual fall (falling ramp). The asynchronous changes consisted of random disturbances to steam flow. The severity of the continuous and asynchronous changes varied as a function of the elapsed time of an experimental trial.

The probes occurred at random times during a trial and were administered when the next screen update was scheduled to appear. An auditory tone sounded, and a description of the probe was presented in a text panel at the top of the screen (e.g., "Enter % value for SF"). The participants entered a numeric response on the keyboard; they were instructed to respond as accurately and as quickly as possible. Accuracy scores were computed by taking the absolute value of the difference between the observer's estimate and the actual value that appeared on the screen. Latency was measured from the time that the prompt appeared until the participant entered the first digit of their response (0.01 s accuracy). Participants were given feedback on accuracy.

Two types of probes were administered: low-level data and high-level property. A low-level data probe required an observer to enter the value for one of the four individual variables; a high-level property probe required participants to enter the absolute value of the difference between either (a) ISGL and CSGL (energy balance) or (b) SF and FF (mass balance). Two types of methodologies were used to administer probes: visual and memory. In the memory methodology, the screen was blanked before the probe was presented; this forced the observer to recall the information from memory. In the visual methodology, the screen was not blanked before the probe (i.e., the display remained available during a probe). An algorithm ensured an approximately equal distribution of probes among the various conditions.

RESULTS

Tests for outliers (accuracy, latency, and time on task) were performed using standardized deviate statistics (Barnett & Lewis, 1984; Lovie, 1986; Ratcliff, 1993). The test is described in Lovie (1986, pp. 55–56): $T_1 = (x_{(n)} - \bar{x})/s$, where $x_{(n)}$ 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. An outlier status for either accuracy or latency resulted in the removal of both scores. Of the 5503 probes that were administered, 131 were discarded (2.38%); Wilcoxon signed ranks tests revealed that the distribution was random. Of the 726 time-on-task scores, 3 were discarded (0.41%). Analyses with and without outliers revealed the same general pattern of significant findings; the minor differences do not change the nature of the conclusions that have been drawn.

Accuracy

Scores were averaged across experimental session, and a 2 (display) × 2 (probe) × 2 (method) repeated-measures analysis of variance (ANOVA) was performed. The main effects of display, $F(1, 7) = 48.16$, $p < .0003$, method, $F(1, 7) = 186.43$, $p < .000003$, and probe, $F(1, 7) = 30.47$, $p < .0009$, and the Display × Method, $F(1, 7) = 11.71$, $p < .02$, Display × Probe, $F(1, 7) = 6.27$, $p < .05$, Method × Probe, $F(1, 7) = 56.04$, $p < .0002$, and Display × Method × Probe, $F(1, 7) = 12.00$, $p < .02$, interactions were significant. The means for the three-way interaction are illustrated in Figure 2. Planned comparisons to assess performance differences between the composite and baseline displays were conducted for each methodology/probe combination. In Figure 2 these comparisons are identified by pairs of filled versus open symbols that have a similar shape and methodology (e.g., the symbols labeled 1 and 2). For the visual methodology, the composite display improved performance significantly for both low-level data (1 vs. 2), $F(1, 7) = 106.58$, $p < .00002$, and high-level

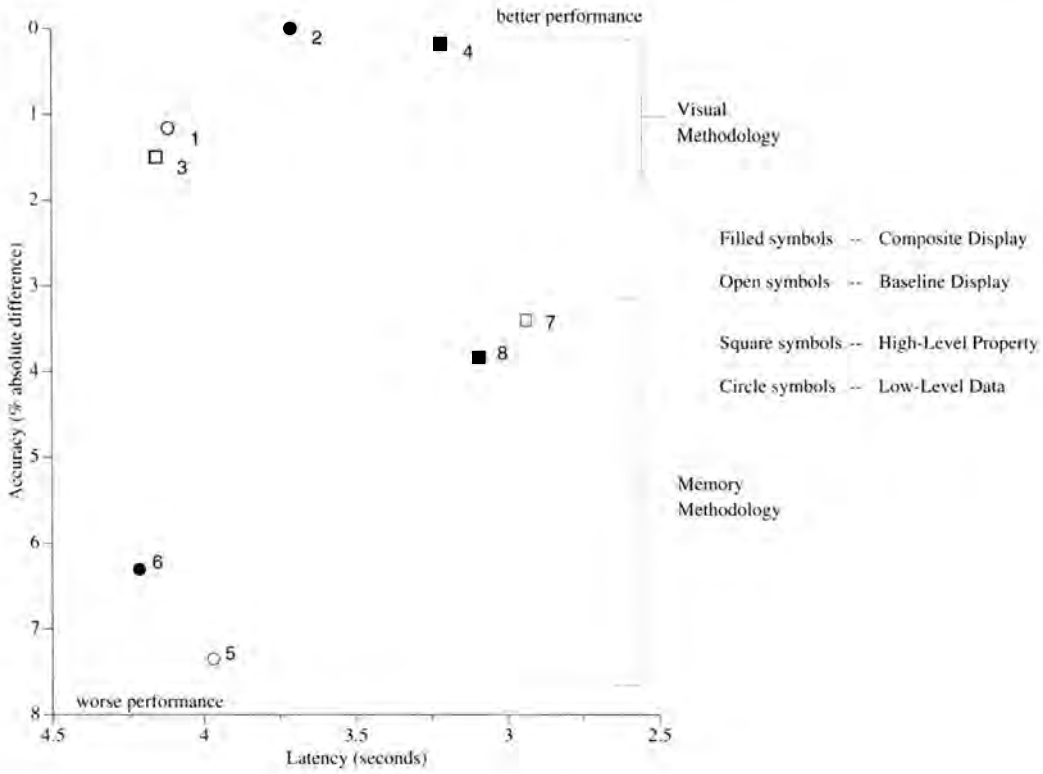


Figure 2. Mean levels of performance for the Display × Probe × Methodology interaction effects. Accuracy is plotted on the *y* axis and latency on the *x* axis. The scales on these axes are reversed so that better performance is located in the upper right portion and worse performance is located in the lower left portion. Means obtained with the visual and memory methodologies are identified with the verbal and graphic labels on the right side of the graph.

property probes (3 vs. 4), $F(1, 7) = 32.41$, $p < .008$. For the memory methodology, the composite display improved performance significantly for low-level data probes (5 vs. 6), $F(1, 7) = 8.38$, $p < .03$, but not for high-level probes (7 vs. 8).

Latency

Similar analyses were performed for latency. The ANOVA revealed that the main effects of display, $F(1, 7) = 7.49$, $p < .03$, and probe, $F(1, 7) = 30.55$, $p < .0009$, and the Display × Method, $F(1, 7) = 15.04$, $p < .007$, Display × Probe, $F(1, 7) = 7.02$, $p < .04$, Method × Probe, $F(1, 7) = 54.30$, $p < .0002$, and Display × Method × Probe, $F(1, 7) = 6.69$, $p < .04$, interaction effects were significant. The planned comparisons for the visual methodology revealed significant differences for both low-level data (1 vs. 2), $F(1, 7) = 11.79$, $p < .02$, and high-level property probes (3 vs. 4), $F(1, 7) = 12.05$, $p < .02$ (better performance for the composite display). For the memory methodology there were significant differences for high-level property probes (7 vs. 8), $F(1, 7) = 18.35$, $p < .004$ (better performance for the baseline display), but no significant differences for low-level data probes (5 vs. 6).

Time on Task

The time-on-task scores were averaged across experimental session, and a 2 (display) × 3 (ramp) × 4 (initial ISGL) repeated-measures ANOVA was performed. The main effects of ramp, $F(2, 14) = 11.68$, $p < .002$, and initial ISGL level, $F(3, 21) = 7.28$, $p < .02$, and the Ramp × ISGL interaction, $F(6, 42) = 4.54$, $p < .02$, were significant. The main effect of display, $F(1, 7) = 1.61$, $p < .25$, failed to reach significance. The significant interaction effect indicated

that specific combinations of ramp and initial starting position were particularly difficult. The poorest performance occurred for a low initial ISGL paired with a rising ramp and for a high initial ISGL paired with a falling ramp.

DISCUSSION

The results obtained with the control methodology (i.e., time on task) revealed no significant differences between displays. In contrast, the results obtained with the memory and visual methodologies revealed numerous significant effects. The Display × Probe × Methodology interaction effects for both accuracy and latency are illustrated in Figure 2. As this figure reveals, the methodology used to evaluate performance had a major impact on the results (the results obtained with each methodology are indicated by the textual labels on the right side of the graph).

The overall accuracy of performance was substantially better with the visual methodology than with the memory methodology. In addition, the pattern of results for display and probe differed between methodologies. The type of display had the most pronounced impact on performance for the visual methodology (filled vs. unfilled symbols); the type of probe had the most pronounced impact for the memory methodology (square vs. circle symbols). These diametrically opposed patterns of results provide very different answers to issues in configural display design. An interpretation requires a consideration of the joint constraints on performance that were introduced by the evaluation goals, methodologies, probe types, and displays.

With respect to the evaluation goals, there are many levels at which interface evaluations can and perhaps should be conducted. In the Rasmussen, Pejtersen, and Goodstein (1994) framework, the visual and memory methodologies are located at the *Boundary 1* level of evaluation. These particular evaluations focus on the physical properties of a display and how these properties relate to and interact with the participants' perceptual/cognitive capabilities. In the process of designing a display, the designer encodes information using particular graphical features. This design has implications for the ease or difficulty with which an observer can decode that information (Cleveland, 1985). The present evaluations focus on how well participants could decode or remember information (i.e., provide a quantitative estimate of current system state) that was encoded into the alternative displays. Although there is a direct link to a domain, the tasks are actually defined by the perceptual characteristics of the display at the instant that the probe occurs. In contrast to other types of tasks (e.g., time on task), there was no requirement to do anything with the obtained information other than to report it.

The primary finding obtained with memory methodology is that there were very large effects for probe type: Participants had greater difficulty completing the low-level data probes than they did completing the high-level property probes. The combination of low-level data probes and memory methodology produced the poorest performance in the experiment, regardless of whether the baseline (open circle – 5) or the composite (filled circle – 6) display was present. In contrast, performance for high-level property probes was substantially better (7 and 8). Although there were significant differences between displays, the patterns did not reveal a consistent advantage for either display type. For the accuracy of low-level data probes, the composite display was superior to the baseline display (6 vs. 5, respectively). For the latency of high-level property probes, the baseline display was superior to the composite display (7 vs. 8, respectively).

Interpreting these results involves a consideration of both the characteristics of the information probe tasks and the perceptual/cognitive capabilities that could be brought to bear for their completion. As discussed previously, the correct response to the probes was defined by the visual characteristics of the display at a particular point in time. However, the memory methodology forced participants to use information in memory to generate an estimate because the display was removed from sight prior to the administration of a probe.

As mentioned in the introduction, there are several reasons for the fact that memory for this particular type of information is severely limited. Veridical, detailed visual information

about the displays would not have persisted long enough to be useful (Sperling, 1960). In addition, the nature of the probe tasks limited the extent to which information in long-term memory could be used. If long-term memory was involved at all, it is likely that it was information in episodic (temporally based, autobiographical information) as opposed to semantic (symbolic, meaningful information) long-term memory (Tulving, 1972) that would have been used. In contrast to semantic memory – which is highly organized, stable, and immense – episodic memory is "in a constant state of change, and information there is often transformed or made unretrievable" (Klatzky, 1980, p. 179). For example, proactive interference (Keppel & Underwood, 1962) might have contributed to degraded performance (i.e., the value of low-level data and high-level properties from previous probes or trials might have interfered with participants' ability to perform subsequent probes).

These observations form the basis for the interpretation of results obtained with the memory methodology. The most likely explanation for the generally lower levels of performance is that the information about the physical details of the display that was available with the memory methodology was simply less detailed than the information available with the visual methodology. A similar explanation could account for the finding that the design techniques were not effective. These techniques were hypothesized to improve performance because they provided visual information relevant to the probes (exact digital values and a visual structure that highlighted relationships between data markers and relevant scales). However, this visual information was ineffective in improving performance because it was not preserved in sufficient detail in the memorial representation.

The interpretation of the results obtained for probe type involve a similar but slightly more complicated line of reasoning. Intuitively, the low-level data probes should be easier to complete than the high-level property probes (only the value of one variable, as opposed to the absolute difference between two variables, needs to be reported). However, the opposite pattern was found with the memory methodology. To understand these results, the constraints of the task cannot be isolated from (i.e., must be considered in conjunction with) the representation that the display provides.

In the case of the low-level data probes, the observer had to provide an estimate of horizontal (for SF or FF) or vertical (for ISGL or CSGL) extent associated with an individual variable (i.e., the distance between the appropriate side of the rectangle and the appropriate axis). To produce this estimate, the observer would need a memory encoding that was sufficiently detailed to allow the recall of some (if not all) of the following information: (a) the shape and size of the rectangle, (b) the position of the rectangle relative to the x and y axes, and (c) the location of the relevant variable in the rectangular form. In contrast, to complete a high-level property probe, the observer needed only a memorial representation that was sufficiently detailed to produce an estimate of the height or width of the rectangle, given that these emergent features corresponded directly to the information requested in this probe (the features of energy balance and mass balance, respectively). If the relationships that the participants had to report were *not* mapped into salient emergent features of the display, then a very different pattern of results might have been obtained. For example, if participants had been asked to report the absolute value of the difference between steam flow (left or right side of the rectangle) and indicated steam generator level (top or bottom of the rectangle), then performance would probably have been as bad as, or worse than, performance for individual variables.

The results obtained with the visual methodology revealed a very different pattern of results. For low-level data, the participants responded significantly more quickly and accurately with the composite display than with the baseline display (1 vs. 2). In addition, there were no design trade-offs with respect to the extraction of higher-level properties. Performance was also significantly faster and significantly more accurate for the composite display than for the baseline display (3 vs. 4). Figure 2 shows that the means obtained with the visual methodology were organized according to the display that participants observed (filled vs. unfilled symbols) rather than to the probe that

they completed (square vs. circle symbols). Thus, applying the complementary design strategies was very effective in improving performance relative to the baseline configural display when the displays were available for inspection.

The interpretation of these results is fairly straightforward. The visual methodology allowed participants to use perceptual systems of virtually unlimited capacity to complete the probes. Therefore, the results provide an assessment of the extent to which individuals could extract or decode the information that had been encoded into the displays. The additional visual information provided by the design strategies, as a whole, improved participants' ability to extract low-level data from configural displays and did not interfere with (and actually improved) their ability to extract high-level properties.

The results of the present experiment highlight the critical role of methodology in the evaluation of configural displays: The two methodologies revealed diametrically opposed patterns of results with regard to important issues in configural display design, even though the same participants, displays, and probes were used. Both sets of results are reliable. The memory methodology results are similar in many respects to those obtained by Bennett et al. (1993), and the visual methodology results were replicated and extended by Bennett and Walters (in press). Other experiments that have used both memory-based and perceptual-based methodologies concurrently in the investigation of either display design or closely related topics have also obtained different patterns between methodologies (Howard & Kerst, 1981; Moyer, Bradley, Sorensen, Whiting, & Mansfield, 1978; Moray et al., 1993; Scott & Wickens, 1983; Wickens, Merwin, & Lin, 1994).

The question for researchers who are investigating issues in configural displays (and for that matter, in interfaces in general) is whether or not one set of results is more meaningful for design and, correspondingly, whether or not one category of methodologies is preferable for future evaluations. The visual methodology probably provides the more representative results, at least under the circumstances and goals of the present evaluation. The critical feature that distinguishes the two methodologies is their reliance on either perceptual or memory processes. The memory methodology provided results that testify to an individual's capability to recall information encoded into a display, and it is clear that performance under this methodology was substantially worse than performance when the display was visible.

Although it is true that individuals might occasionally be required to rely on memory during interaction, a primary goal of effective interface design should be to minimize these instances (consistent with the results obtained in this study). Thus the visual methodology, which provides an assessment of individuals' ability to extract information that has been encoded into a display, is preferable to the memory methodology. The visual methodology will inform researchers about factors in design that influence performance when higher-level interface design goals have been met (i.e., when the displays that are relevant to a decision are visible).

From a design perspective, the results obtained with the visual methodology represent progress toward a goal: the design of graphical formats that are capable of supporting effective information extraction along a continuum ranging from low-level data to high-level properties. One goal of the present research was to investigate the utility of several design strategies in improving the capability of configural displays to support performance at tasks that require the consideration of low-level data. When considered as a whole, the design strategies were clearly effective in doing so. Bennett and Walters (in press) extended and clarified these results by assessing the contribution of each design strategy (alone and in various combinations) to improved performance for the extraction of low-level data. The presence of digital values was particularly effective, but all design techniques (with the exception of color coding/layering/separation) produced significantly better performance. Whether or not these techniques will improve performance with configural displays relative to alternative formats (e.g., bar graphs) remains an open empirical question. However, we believe that this is likely to be the case.

From the perspective of evaluation, one final question must be considered: Are there any circumstances under which memory-based methodologies might prove to be useful in interface and display evaluation? The answer is a qualified "yes." These methodologies can be useful when meaningful, semantic information is required to complete experimental tasks (in contrast with the present experiment). For example, Vicente and his colleagues (Moray et al., 1993; Vicente, 1988, 1992) have proposed an adaptation of the memory recall paradigm for display evaluation. In this procedure domain information is presented using alternative display formats, and the capability of expert users to recall this information is assessed at a subsequent point in time. The level of recall should be determined by the extent to which a display representation allows users to chunk relevant information about the domain and thereby improve recall performance.

Another example can be found in more traditional human-computer interface design, which has increasingly involved the use of metaphors (e.g., the desktop metaphor) or visual icons. These displays (metaphors and icons) must be assessed for their consistency with the conventional, general knowledge of the target population of users. For example, Rasmussen et al. (1994) used visual and verbal tests of associative memory to evaluate icons for use in a library information retrieval system. In both of the examples, memory methodologies have been used to assess the extent to which features in display design are consistent with the detailed, semantic knowledge structures of the associated users.

ACKNOWLEDGMENTS

The authors would like to thank David Woods, William Howell, John Flach, and three anonymous reviewers for discussions and comments on earlier drafts. Funding was provided by Wright State University (Research Challenge and Research Incentive Grants). The part-task simulation was originally developed at the Westinghouse Electric Corporation, primarily in the Department of Human Sciences Research, with the assistance of several individuals and organizations.

REFERENCES

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data.* Chichester, NY: Wiley.

Barnett, B. J., & Wickens, C. D. (1988). Display proximity in multicue information integration: The benefits of boxes. *Human Factors, 30,* 15–24.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

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

Bennett, K. B., & Walters, B. (in press). Configural Display Design Techniques Considered at Multiple Levels of Evaluation. *Human Factors.*

Bransford, J. D., & Franks, J. J. (1972). The abstraction of linguistic ideas: A review. *Cognition, 1,* 211–250.

Cleveland, W. S. (1985). *The elements of graphing data.* Belmont, CA: Wadsworth.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229,* 828–833.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Gillan, D. J., & Lewis, R. (1994). A componential model of human interaction with graphs: I. Linear regression modeling. *Human Factors, 36,* 419–440.

Gillan, D. J., & Richman, E. H. (1994). Minimalism and the syntax of graphs. *Human Factors, 36,* 619–644.

Howard, I. H., & Kerst, S. M. (1981). Memory and perception of cartographic information for familiar and unfamiliar environments. *Human Factors, 23,* 495–504.

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior, 1,* 153–161.

Klatzky, R. L. (1980). *Human memory: Structures and processes.* San Francisco: Freeman.

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.

Meyer, J., Shinar, D., & Leiser, D. (1997). Multiple factors that determine performance with tables and graphs. *Human Factors, 39,* 268–286.

Moray, N., Jones, B. J., Rasmussen, J., Lee, J. D., Vicente, K. J., Brock, R., & Djemil, T. (1993). *A performance indicator of the effectiveness of human-machine interfaces for nuclear power plants* (Report No. NUREG/CR-5977). Washington, DC: U.S. Nuclear Regulatory Commission.

Moyer, R. S., Bradley, D. R., Sorensen, M. H., Whiting, J. C., & Mansfield, D. P. (1978). Psychophysical functions for perceived and remembered size. *Science, 200,* 330–332.

Parkin, A. J. (1984). Levels of processing, context, and facilitation of pronunciation. *Acta Psychologica, 55,* 19–29.

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

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

Scott, B., & Wickens, C. D. (1983). Spatial and verbal displays in a C3 probablistic information integration task. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 355–358). Santa Monica, CA: Human Factors and Ergonomics Society.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*(Whole No. 498).

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic.

Vicente, K. J. (1988). Adapting the memory recall paradigm to evaluate interfaces. *Acta Psychologica, 69,* 249–278.
Vicente, K. J. (1992). Memory recall in a process control system: A measure of expertise and display effectiveness. *Memory and Cognition, 20,* 356–373.
Wickens, C. D., & Andre, A. D. (1990). Proximity compatibility and information display: Effects of color, space, and objectness on information integration. *Human Factors, 32,* 61–78.
Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37,* 473–494.
Wickens, C. D., Merwin, D. H., & Lin, E. L. (1994). Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh. *Human Factors, 36,* 44–61.
Woods, D. D., Wise, J. A., & Hanes, L. F. (1981). An evaluation of nuclear power plant safety parameter display systems. In *Proceedings of the Human Factors Society 25th Annual Meeting* (pp. 110–114). Santa Monica, CA: Human Factors and Ergonomics Society.

Kevin B. Bennett received his Ph.D. in applied experimental psychology in 1984 at the Catholic University of America. He is an associate professor in the Psychology Department at Wright State University.

Michael Payne received his M.S. in human factors psychology in 1999 at Wright State University. He is a human factors engineer in the User Centered Design Group at Intel Corporation, Hillsboro, Oregon.

Jeffrey Calcaterra received his B.A. in psychology in 1994 at North Carolina State University. He is a human factors engineer at IBM Corporation in Research Triangle Park, North Carolina.

Bob Nittoli received his M.S. in human factors psychology in 1991 at Wright State University. He is a usability engineer at eLink Commerce, Glendale, California.

Date received: April 13, 1998
Date accepted: August 11, 1999