ARI TECHNICAL REPORT
TR-77-A7

TASK AND OBSERVER SKILL FACTORS IN ACCURACY OF ASSESSMENT OF PERFORMANCE

by

Darren Newtson
University of Virginia

APRIL 1977

Grant No. DAHC 19-74-G-0016
ARI Themes Program

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
1300 Wilson Boulevard
Arlington, Virginia 22209

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

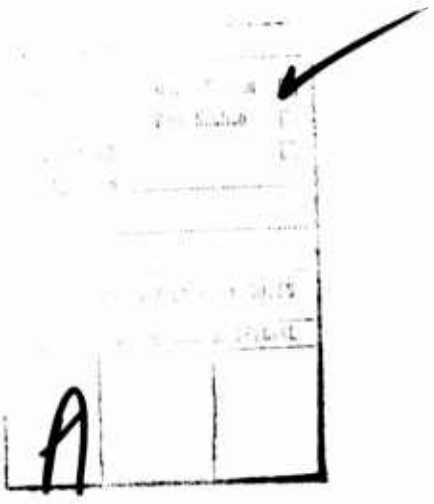
A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

J. E. UHLANER
Technical Director

W. C. MAUS
COL, GS
Commander

Research accomplished
under contract to the Department of the Army

University of Virginia



NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 1300 Wilson Boulevard, Arlington, Virginia 22209.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report TR-77-A7	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: TASK AND OBSERVER SKILL FACTORS IN ACCURACY OF ASSESSMENT OF PERFORMANCE	5. TYPE OF REPORT & PERIOD COVERED: Final Report 5/1/74 to 10/31/76
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Darren Newtson	8. CONTRACT OR GRANT NUMBER(s): DAHC 19-74-G-0016
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Virginia, Charlottesville, Virginia 22901	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q161102B74F
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences PERI-11, 1300 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: April 1977
	13. NUMBER OF PAGES: 80
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

Research was performed under the ARI Themes program; Dr. Milton H. Maier of the ARI IndividualTraining and Skill Evaluation Technical Area, was technical monitor.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Observer Skill, Performance Assessment, Behavior Perception, Observer Accuracy

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A proposed measure of the perceptual organization of ongoing behavior was applied to the problem of operationalizing observer skills. Twelve experimental studies were completed. Evidence was obtained verifying that the proposed measure taps a low-level perceptual process, the subjective organization of action, and with a high degree of reliability. Further studies established that the resulting action units are true phenomenal partitions of observer experience, achieved by the discrimination of successive points of definition in the behavior stream.

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

(20 continued)

On the basis of these data, a conceptual model was proposed with specific, testable implications for the study of observer skill. Finally, the applicability of the model to accuracy of observer judgement was verified, and evidence was presented establishing that two independent components of observer accuracy exist; consistent with the proposed conceptual model, it was determined that differences in observer accuracy are due in part to differences in the skill of the observer in the perceptual organization of task performance.

Abstract

A proposed measure of the perceptual organization of ongoing behavior was applied to the problem of operationalizing observer skills. Twelve experimental studies were completed. Evidence was obtained verifying that the measure taps a low-level perceptual-attentive process, the subjective organization of action, and with a high degree of reliability. Further studies established that the resulting action units are true phenomenal partitions of observer experience, achieved by the discrimination of successive points of definition in the behavior stream. On the basis of these data, a conceptual model of observation was proposed with specific, testable implications for the study of observer skill.

Finally, the application of the model to accuracy of observer judgement was verified, and evidence was obtained establishing that two independent components of observer accuracy exist, stereotypic and differential accuracy. The former refers to the ability of observers to judge the absolute skill level of the group of stimulus persons judged; the latter refers to observers' ability to correctly rank stimulus persons on the skill dimension. Consistent with the proposed conceptual model, it was determined that differences in observer accuracy are due in part to differences in the skill of the observer in the perceptual organization of task performance. In general, it appears that skilled perception of performance is a necessary, but not completely sufficient, condition for accuracy of observer judgement.

Table of Contents

With the development of precise descriptions of the component tasks for specific job classifications, and the adoption of competence in these tasks as a primary goal of training, the need for more precise and adequate means of evaluation of task performance has become increasingly important (Maier, 1976). Due to the nature of many important task performances, furthermore, the use of objective performance tests is limited. Competence in many job classifications depends upon the ability of the trainee to perform a series of organized, goal-directed actions in an efficient and coordinated manner. When such tasks are embedded in an overall team effort, or they are specific parts of a larger task organization, objective performance indices may be both costly to obtain and of questionable validity. In these applications the least costly and most efficient means of performance assessment may be to employ skilled observers to evaluate the competence of trainees.

The value of observer ratings of performance, however, depends upon the validity of those ratings. Not all observers may be equally competent, reliable or skilled in the evaluation of a given task performance. When observers disagree, moreover, it is important to locate the nature and source of the disagreement. There would seem to be, on the face of it, two different sources of information one could turn to for means of dealing with these issues.

Given the importance of behavior observation in a wide range of academic disciplines, one might expect to find a fully developed research literature on the nature and limitations of this methodological technique. Methodological investigations in this area, however, have focused primarily on the development of sampling techniques to insure unbiased estimates of the frequency of specified behaviors (cf. Altman, 1974), and conditions under which reliabilities of observers in applying a priori coding schemes to particular classes of behavior may be maintained. As a result, this literature is surprisingly uninformative with respect to the question of what observers know, and how they come to know it; instead, it has focused upon the validity of particular analytic coding schemes, and upon the reliability of observers in applying those coding schemes.

The irony of this neglect comes from the fact that the proponents of observational methods, while decrying the artificial, analytic constraints of laboratory research as ecologically invalid, have devoted themselves to bringing equally artificial analytic schemes to ongoing behavior in natural settings. Barker (1963) coined the term "behavior units" for the naturally-occurring organizations in ongoing behavior, differentiating them from that he termed "behavior tesserae." Behavior tesserae are, "...fragments

of behavior that are created or selected by the investigator in accordance with his scientific aims (p. 2)." As Barker (1963) notes, while investigators have freely extracted such fragments according to their own preconceptions and analytic intentions, little attention has been paid to identifying natural, ecologically valid units of behavior.

Barker (1963) suggests that one reason for this neglect is the widespread view that, "...the course of behavior is such a complicated, unstable phenomenon that it is not amenable to ordering in lawful ways (p. 6)." Accordingly, students of human behavior have relied upon what Barker terms "structure-destroying or structure-ignoring research methods," and have largely neglected the development of "tender, sensitive, non-destructive techniques for exploring the natural units of their phenomena (pp. 2-3)." The development of such techniques, he argues, requires a pluralistic, open-minded, empirical, proto-theoretical approach, "where investigation must follow the canons of discovery rather than those of scientific verification (p. 10)."

The relevance of Barker's (1963) criticisms of analytic methodologies to the present discussion follows from the necessity, if one is to understand the nature and limitations of observer skill in performance evaluation settings, of understanding the phenomenal basis of observer judgements. That is, observers do not base their judgements of an ongoing task performance on the behavior per se, but upon their subjective perceptual organization of that behavior. As Barker (1963) and Barker and Wright (1955) point out, behavior as a stimulus contains a wide range of real, natural organizations, from brief, fleeting reflexes to persistent, goal-directed activities lasting for periods of years. Only a limited range of these organizations in ongoing behavior are susceptible to normal observation. Barker and Wright (1955) draw an analogy between perceived units of action and objects that can be seen with the naked eye (p. 6). As some objects are so small as to be below the limits of visual acuity, so some behavioral organizations are so brief as to pass unseen by human observers. Similarly, as some objects are so large as to defy normal observation (e.g. an entire mountain range), so some behavior organizations are so long in duration as to escape comprehension. While such organizations may in fact exist in the behavior stream, they are not normally accessible to human observation. As Barker and Wright (1955) pointed out, we do not live in a phenomenal world, "...full of muscle twitches and lifetime undertakings (p. 245)." They termed this range of experience of behavior organization the "normal behavior perspective."

Barker and Wright (1955) were concerned with justifying the perceived units of action that could be reliably discriminated by observers as valid data for the objective analysis of behavior organization. Their approach was founded on the assumption that, as persons are generally capable of correctly interpreting and responding to the behavior of others, the organizations they interpret and respond to must have some valid basis in the behavior of other persons. Their goal, then, was to start with the analysis of intuitive behavior organization, lifting themselves by empirical "bootstraps" to a point where the inherent, natural units of behavior organization could be more clearly specified.

As noted, our concern with respect to observer skill is more with the processes of subjective behavior organization than with the ultimate reality of the structures that observers discriminate. Insofar as observers must judge the skill level of ongoing task performances from those performances, the processes of behavior observation must begin with the perception of that performance.

A second possible source of information as to the nature and components of observational processes, therefore, is research on visual perception itself. Given that a task performance is an ongoing event - a behavioral event - we might expect to find useful evidence as to the processes by which ongoing behavior as a stimulus is perceptually organized into a series of temporally connected organized actions. Research on visual perception, however, has failed to address these issues. Neisser (1975) has forcefully criticized research in perception and information processing as failing to deal with perception as an ongoing process, as, in the perceptual organization of ongoing behavior, it must be. Neisser (1975) notes that a consequence of this failure is a "tachistoscopic" view of perception as beginning with sensory stimulation on the surface of the retina and ending with a percept in the mind. Such an approach has ignored processes of active information search and anticipation that are essential to the functioning of perceptual systems in a normally information-rich, ongoing stimulus environment. Jenkins, Wald, & Pittenger (in press) offer a similar criticism of current perceptual theory, as does Miller and Johnson-Laird (1976). All of these authors concur in their assertion that the perceptual organization of events is a basic, neglected problem in visual perception.

By ongoing event perception, it should be pointed out, we do not mean perception of movement. Miller and Johnson-Laird (1976) note that the two are often equated, but argue that a strong distinction must necessarily be made between them (p. 85). The difference may perhaps best be appreciated in terms of an analogy

between words and behavioral events, or actions, and between sound and movement. Words are composed of sounds, as actions are composed of movement. As some sounds are not words, so some movements are not actions. Similarly, as some dimensions of sound in words are readily discriminable (e.g., pitch, rate, accent, etc.), yet do not define the meaning of words, so some aspects of movement in actions may also be readily discriminable, without being basic to the comprehension of actions. Assertions as to the stimulus bases of perceived actions, therefore, require empirical support, and the assumption of a simple isomorphism between the stimulus and its phenomenal apprehension may be in error.

A third possible approach to the problem of specifying observer skill might be to assume that the perceptual organization of a performance is wholly determined by stimulus factors, and hence is relatively constant across observers. Differences between observers in judgemental accuracy would then be sought in the judgemental process, rather than in the initial perceptual selection of information from the ongoing event. This kind of "social judgement" approach has been taken to questions of accuracy of clinical judgements (Sarbin, Taft, & Bailey, 1960; Bieri, Atkins, Briar, Leaman, Miller, & Tripodi, 1966). Such an approach, however, commits one to a possibly untenable assumption. Wiggins (1969) investigated judgements of intelligence of stimulus persons, and concluded that the single most important predictor of judgemental accuracy was the predictive validity of the cues selected as a basis for the judgement. Given that the most predictive set of cues was selected, differences in cue weighting - presumably a function of the judgemental process - produced only very slight differences in judgemental accuracy. This study, it should be noted, is one of very few that investigated cue selection in social judgement.

Further evidence bearing on this assumption is provided by a series of studies reported by Newtson (1973). Newtson (1973) noted that current formulations in social perception have assumed that the perceptual organization of observed behavior is constant across observers, and proposed a means of testing this assumption. Central to Newtson's (1973) test of the hypothesis that variation in perceptual organization would affect outcomes of social judgement was the proposal of a technique for measuring the subjective unit of perception of action. A careful search of the literature provided only two previous attempts to measure this phenomenon, both seriously flawed.

Lyons (1956) attempted to measure the unit of perception by verbal report. Subjects viewed two problem-solving sequences under instructions to describe into a tape recorder "all the different

things" the actor did. These reports were scored by judges for reference to molar vs. moleculor units of behavior. One of the problem-solving sequences was "apparently random," consisting of trial-and-error attempts to solve an insight problem; in the second sequence, the actions were immediately and consistently directed towards a clearly evident goal (pp. 48-49). Lyons (1956) compared live presentation to films, and found no differences between the two. Subjects in his experiment were matched groups of schizophrenics and normals. Schizophrenics were found to employ molecular units for both sequences. Normals, however, used smaller units for the "apparently random" sequence, but shifted to larger, more molar units for the goal-directed sequence.

As a measurement technique, this procedure has serious shortcomings. It requires the assumption that the units of behavior perception are capable of direct verbal expression, and that dimensions of variation of perceptual units will be reflected directly in verbal reports. In addition, it is extremely difficult to estimate the degree of consensus among perceivers.

Dickman (1963) improved the precision of unit measurement, but at the cost of immediacy. Dickman had subjects view a film sequence, and then sort a sequence of cards with the behavior in the film written on them. Each card corresponded to a "minimal molar unit," an action so small that further breakdown would result in descriptions of muscle movements. The cards were presented in a numbered sequence so that they described the behavior presented in the film. Subjects were instructed to divide the cards into groups so that each group represented a "happening" in the film in a way that seemed most natural to them. Dickman reported that subjects readily understood his instructions and performed the task without difficulty. Results indicated better than chance agreement on 1) the general patterning of points of division and continuity; and 2) the designation of "break" and "continue" at one-half of the individual choice points.

Dickman (1963) viewed his results as somewhat paradoxical: there was high agreement at some points, and high variability at others. Closer examination of the data indicated that the variability was due to differing sizes of behavior units, varying over a hierarchical structure of goals and subgoals.

Newtson (1973) reasoned that the approaches of Lyons (1956) and of Dickman (1963) could be combined to yield an adequate technique for the measurement of the unit of perception, and hence be used to test the assumption of constant perceptual input in current social judgement theories. The technique consists of

providing subjects with a button operating a continuous event recorder, and instructing them to press the button when, in their judgement, one meaningful action ends and a different one begins. Measurement is thus immediate, as in Lyons (1956), yet precise estimates of agreement may be obtained, as in Dickman (1963). Employing this methodology, Newtson (1973) conducted two experiments.

The first experiment directly manipulated the size of the unit of perception by instruction. Drawing on concepts from information theory, it was predicted that subjects using more units to perceive a given sequence of behavior should be in a higher information state (cf. Kelley, 1967) about the actor than subjects using fewer, larger units of perception. In this experiment, subjects were instructed to mark off either 1) the largest, or 2) the smallest, actions that seemed natural and meaningful to them. Results of this experiment were perfectly in accord with predictions: subjects employing smaller, and hence more, units of perception were more confident in their judgements of the person, made more differentiated judgements, and made more dispositional attributions for his behavior.

In a second study, unitization was employed as a dependent measure. Subjects viewed one of two videotapes of an actor performing a molecule-model assembly task. In one of the tapes, a 30-second insertion of an unexpected action was made. It was predicted that, to the extent that variation in perceptual organization performs an information regulation function, subjects viewing the unexpected action would subsequently employ finer units of perception as a means of increasing their understanding of the event. This prediction was also confirmed.

The results of these investigations demonstrate quite clearly that the perceptual organization of an ongoing performance may vary, and that that variation may systematically alter judgements based upon that observed performance. It is possible, therefore, that differences in observer skill might be directly reflected in differences in perceptual segmentation of relevant performances. Confirmation of this proposition, furthermore, could have wide implications for the nature of observer skill, as well as considerable value in application. If it can be demonstrated that the perceptual segmentation of a sequence is a prime determinant of the judge's information base, we will have succeeded in operationalizing observer skill. This demonstration depends, however, upon evidence that the units discriminated with the unit marking procedure are low-level, initial perceptual units, as opposed to higher-level, rationalistic discriminations. That is, we must have evidence that we are indeed tapping the perceptual information base of observation directly, as opposed to higher level, correlated

epi-phenomena of observational processes. Evidence that this is the case may be provided from three kinds of sources: 1) from evidence that the processes involved are indeed low-level processes; 2) from evidence that the perceptual units discriminated by the procedure are phenomenal wholes in the experience of the observer; and 3) from evidence that such units are based upon a physically present stimulus property. Concurrently, given evidence as to the underlying properties of behavior units, we may, as Barker (1963) suggests, lift ourselves by empirical bootstraps to a point where we may formulate consistent, testable hypotheses as to the nature of observational processes and their dimensions and limitations. Our goal, then, is the development of a general, theoretical model of the observer with specific implications for the nature of observer skill.

Phase I: The Level of Behavior Segmentation Processes

The logic of the first phase of the research was as follows: In order to identify the level of processing at which unit formation occurs, we may investigate the effects of variables on behavior segmentation previously demonstrated to have effects on both higher and lower-level processes. Given evidence as to the effects of both types of variables, we may then attempt more specific hypotheses as to the nature of the process of behavior observation.

Two variables were investigated: 1) the hedonic relevance of the behavior of the actor for the observer; and 2) physiological arousal during observation.

Experiment One

Jones and Davis (1965) argue that behavior which has motivational significance for the perceiver may alter his interpretation of that behavior. They cite a number of studies indicating that behavior which bears on the receipt of rewards for the observer leads to increased "correspondence" of inference. "Correspondence" refers to the information value of a given action or actions to the perceiver (Jones and Davis, 1965; p. 264); operationally, correspondence means ratings toward the extremes of impression dimensions given with confidence. All of the studies cited, however, rely upon presentations of the behavior of the target person in written form. It has not been established whether these effects occur in direct behavior observation, or, if they do, whether the effect is restricted to inference based upon those actions, or alters the interpretations and perceptual organization itself.

With respect to the question of the level of unitization processes, we might expect to find a close relationship between inference processes and unitization to the extent that our measure is directly related to these higher level processes. Thus, one straightforward prediction would be that the higher the hedonic value of the actor's behavior the finer the level of segmentation and the more correspondent the inference.

An experimental study was thus conducted. Subjects segmented a videotape of problem-solving behavior under one of three conditions: High Utility, Low Utility, and Control. Utility of behavior for the observer was varied by varying the amount of money the subject was to receive if the actor succeeded at his task within an alotted time period.

Method

Subjects and Design

Sixty male subjects, both paid and volunteer, were recruited from the Charlottesville area. The thirty-five volunteer subjects were obtained from the introductory psychology courses at the University of Virginia. The remaining twenty-five subjects were paid $1.00 for their participation.

Subjects were evenly distributed in three conditions which varied the utility of the information contained in the videotaped sequence. Utility was manipulated by altering the amount of money subjects stood to win during the course of the experiment in addition to any monies they were paid for participating. These conditions were High Utility ($2.00), Low Utility (25¢), and No Utility (0).

Subjects were assigned to a condition in order of their arrival and were run individually.

Apparatus and Stimuli

The behavior sample of interest and warm-up tape were presented via videotape on a 23-inch television monitor. The sample of interest consisted of a 4½-minute sequence showing a male actor building a tinkertoy structure. Sound was provided. The thirty-second warm-up showed a man playing with a tennis ball.

Subjects recorded judgements of units by pressing a button connected to a continuous event recorder in the next room.

Procedure

Subjects were seated at a table facing the monitor, with the CER button immediately before them, and instructed as follows: "I am going to show you a short videotape of a subject from a study of problem-solving behavior that we did last semester. In that study, subjects were instructed to build a structure out of tinkertoys higher than a line on the wall about four feet high. This sample structure should give you a better idea of what we asked subjects to do. To succeed at the task, they had to complete the construction within three minutes. A tone sounded at exactly three minutes to inform the subjects that his time was up. If they did not make it within three minutes, however, they were told to continue until they did complete the construction. Now to give me a better idea of what happened in that study, I am having five people rate each of the tapes. You will be one of the five raters for the tape that you will see. What I am interested in here is the way people organize their behavior when solving a problem. What I want you to do is mark for me the naturally occuring meaningful actions in the sequence. That is, I want you to press this button firmly, when in your judgement, one meaningful action ends and a different one begins; that is, when the subject stops or finishes doing one thing and starts to do something different. For example, if you observed a person lighting a cigarette and then return the matches to his pocket, you might press the button at the point where he stopped lighting the cigarette and started to put the matches away. You will be given a brief sequence to practice with before you judge the tape we're interested in. Now, there are no right or wrong ways to do this. We just want to know how you do it, and your judgements will be averaged with the four other judges to give us an estimate of the degree of organization of behavior of this subject."

Following the warm-up tape all subjects were instructed: "Now, I don't know exactly which tape you will see. They are just in the order on the tape that we ran them but only a third of our subjects completed the construction in the alotted time and these winners are randomly distributed throughout the tape."

At this point, instructions for the High and Low Utility manipulations were inserted: "Now, as additional reward for helping us, we want to give you some money*, but we couldn't afford very much, so we decided to do this: since only about a third of the subjects completed the problem within the three minute period, we

*At this point paid subjects were reminded that this money would be in addition to the $1.00 they were paid for participation in the experiment.

decided to give those judges who drew a successful subject 25¢ (in the Low Utility group) or $2.00 (in the High Utility group).

At this point all subjects were reminded of the three minute tone which signals whether the subject completed the task in the allotted time. Subjects were instructed to continue marking actions beyond that tone if the subject had not completed the task.

All subjects viewed the same videotaped behavior after which they were asked to complete a task evaluation and Impressions and Attribution measure taken from Newtson (1973). This measure asked subjects to rate the actor on social, intellectual, and evaluative dimensions. After completion of the questionnaire subjects in the High and Low Utility groups were assessed for suspicion of deception, thoroughly debriefed and dismissed. Two subjects were eliminated from the analyses because of their suspicion of the experimental hypotheses.

Results

Mean number of units employed in the No Utility condition were 20.53; in the Low Utility condition, 22.22; and in the High Utility condition, 12.17. High Utility condition subjects used significantly fewer units than subjects in the other two conditions ($p < .05$).

Contrary to the Jones and Davis (1965) hypothesis, no effect of hedonic relevance was observed either on impressions of the person or confidence in impressions. Condition differences were obtained, however, on subjects' estimates of how well the person they observed would perform on subsequent tasks ($F = 7.49$, 2/52 df, $p < .05$). Mean ratings of future performance were (on a nine-point scale ranging from 1 = poorly to 9 = excellently) for the No Utility condition, 4.84; for the Low Utility condition, 5.78; and for the High Utility condition, 4.56. That is, subjects with a small investment in the stimulus person's performance tended to rate him more highly despite his failure; a large investment in his performance, however, produced a substantially lower rating when he failed.

Results of this exploratory study were thus equivocal. Clearly, they demonstrate the hazards of generalizing from paper and pencil social judgement experiments to effects occurring in ongoing observation. As the effects upon both unitization and on estimates of future performance demonstrate, the manipulations did have a differential impact upon the subjects. Contrary to what one might have predicted from Jones and Davis' (1965) hedonic relevance

hypothesis, High Utility condition subjects analyzed the performance less finely than subjects in the other two conditions. At the least then, these data suggest that the relation between unitization and inference processes is an indirect one.

Experiment Two

A second study was conducted to investigate the role of lower level cognitive mechanisms in behavior segmentation. Subjects under the High Utility manipulation in the previous experiment could have decreased unitization because of additional cognitive work (screening irrelevant information, or think about what the most effective procedure might be, etc.), or they could have simply been more aroused. Substantial evidence exists that arousal functions to reduce the range of cue utilization in perceptual tasks (cf. Easterbrook, 1959; Kahneman, 1973; Leventhal, 1970).

In this experiment, subjects segmented two sequences while 1) performing a cognitive interference task, or 2) subject to intermittent white noise, or 3) without interference. In addition to measures employed in the previous study, a recall measure for the sequence was included.

Intermittent white noise was selected as a stressor due to findings that it reliably and effectively induces high states of arousal (Glass and Singer, 1972).

Method

Subjects

Subjects (thirty-six males and nineteen females) were recruited from introductory psychology courses at the University of Virginia, receiving course credit for participation.

Apparatus

A videotape recorder was used to present the behavior sample. Subjects viewed the tape on a television monitor. Subjects recorded judgements of units with a continuous event recorder (CER). An intercom was used to monitor counting in the interference condition.

Stimuli

The behavior sample consisted of two four minute videotapes. The actress in one tape was an undergraduate female and the actor in the second tape was a graduate male. Molly, the actress, played

three games on the videotape. In the first game, the "shuttle-run," there were two lines on the floor, several feet apart. The task consisted of running up to the lines, putting an eraser on each of the lines separately, running back and picking them up. In the second game, the "ball-in-spoon" task, Molly had to pick up a rubber ball out of a small cup with a spoon and put it in a game box lid on the floor. After doing this twice, she returned each rubber ball to the cup with the spoon. In the third game, Molly had to move a square rubber sponge with a stick from ring #1 to #2 to #3 on the floor. Michael, the actor in the second tape, built a wooden tower out of tinkertoys which had to be as tall as a line on the wall.

Procedure

Subjects were seated at a table facing the TV monitor, with the CER immediately before them. The experimenter instructed all subjects as follows:

"In this experiment I am going to show you a videotape of a person playing several games. What I am interested in here are the ways in which people organize or break up another person's behavior. By that I mean that people may break up another person's behavior in different ways. For example, I might turn, walk over, push the door closed, turn, and walk back, and you might see each of those actions as discrete, meaningful act. Or, you might see them as just one action, such as closing the door. What I would like for you to do is mark off for me the naturally-occurring, meaningful actions in the sequence you will shortly see, as you see them. That is, I want you to press this button firmly (indicate CER) when, in your judgement, one meaningful action ends and a different one begins."

"Sometimes, as you watch someone's actions, what you have been seeing as smaller actions start to fit together into larger ones; sometimes the opposite occurs; you start to see the smaller components of larger actions. If, during the course of viewing the tape you find that such a change is occurring for you, feel free to change the size of the units (actions) you have been marking."

"Let me emphasize that there are no right or wrong ways to do this. I just want to see how you do it."

The Control group was then asked if there were any questions.

Other subjects were given either the Noise condition or the Interference condition. Conditions were run alternately, one subject in the Control condition, then one in the Noise condition,

and then one in the Interference condition. Noise condition instructions were:

"Before we begin let me add that you will hear some noise coming from the television monitor, but it is not part of the videotape. The noise is from a white noise generator which we are using to block out background sounds which invariably can be heard coming from other experiments being conducted on this corridor. Are there any questions? O.K., let's begin."

Interference condition instructions were:

"Sometimes people are expected to observe another person's behavior while at the same time doing something else. Therefore, beginning with the first time you press the button, count out loud backwards from 100 (100, 99, 98, etc.), while simultaneously marking off what you consider to be the meaningful actions. Continue to count out loud for as long as the picture appears on the screen. Should you reach zero before the tape has finished, return to 100 and begin again. It is very important that you be as accurate as you can in this counting task, so be sure to keep track of your counting to the best of your ability. However, it is also important that you continue marking the meaningful actions of the behavior sequence. This intercom in front of you will enable me to monitor your counting task from the next room. Do you have any questions? O.K., let's begin.

All subjects were then told by the experimenter:

"Now I'm going to go into the next room to turn the tape on. In order to help standardize the procedure please press the button three times when the person first appears on the screen. This will inform me that the television monitor is working properly and that the tape has begun. Similarly, press the button three times again when the tape ends to let me know that you are finished."

Twenty-nine subjects 1) saw the Molly tape, 2) were given two questionnaires separately, 3) saw the construction tape, 4) then took a recall test on both tapes. Twenty-six subjects 1) saw the construction tape, 2) then saw the Molly tape, 3) then were given the questionnaires and the recall test separately.

Dependent Measures

The dependent measure consisted of a questionnaire asking the subject to rate Molly on nine pairs of traits. Below each pair, subjects were asked to record their confidence in that rating

on a nine-point scale. The nine pairs of traits were obtained from a study of personality inference processes by Barresi (1971). Following the trait and confidence ratings, subjects responded to four items asking them to imagine the person they had observed performing some action (e.g., Dave failed to solve the arithmetic problem). Subjects were then asked to make a forced choice between two explanations of the one they consider "most likely": one an attr bution to an internal, or dispositional, cause (e.g., Dave is poor at arithmetic), the other an attribution to some external property of the situation (e.g., The problem was a very hard one).

Another questionnaire consisted of six questions, asking the person to rate the tasks performed by Molly and how well she did them. A nine-point confidence scale also followed each question on this measure.

The third dependent measure consisted of 18 multiple-choice recall questions to see how well subjects observed the events on both videotapes. The questions ranged from very easy (e.g., Before beginning each activity, Molly (a) wrote something down, (b) read something from a piece of paper, (c) faced the camera, (d) walked around the room) to extremely difficult (e.g., Michael used his (a) right hand to put pieces together and left hand as a steadying influence, (b) right hand to put pieces together and held his left hand at his side, (c) left hand to put pieces together and his right hand as a steadying influence, (d) left hand to put pieces together and held his right hand at his side).

Results

Analysis of variance on the unit measure indicated that only the Arousal condition tended to differ from the Control. Mean number of units were, for the Control, 44.96; for the Arousal condition, 35.14; and for the Cognitive Interference condition, 42.05. The Arousal condition was marginally significant from the other two ($p < .10$, two-tailed). There was extreme heterogeneity of variance in the three conditions.

Results on the recall measure were less ambiguous. The recall test was difficult, as indicated by a mean of only 65 percent correct in the Control condition. Performance in the Arousal condition was marginally poorer ($\bar{X} = 61.6$ percent, $t = 1.42$, $df = 54$, $p < .10$, two tailed), while performance in the Cognitive Interference condition was significantly worse than in both the Control ($\bar{X} = 56.1$ percent, $t = 3.28$, $df = 54$, $p < .05$) and in the Arousal conditions ($t = 1.85$, $df = 54$, $p < .05$).

Discussion

These results provide suggestive evidence on two points. First, they suggest that unit formation occurs at a very early stage in the perceptual interpretation of behavior. That is, the cognitive interference task apparently disrupted memory encoding, without substantially disrupting behavior segmentation. Results in the arousal condition could plausibly be interpreted as reflecting interference at the stage of unit formation, in that fewer units were recorded, and a decrement in recall was observed. That this decrement was less than observed under cognitive interference could be due to the fact that, while information intake was reduced, its subsequent cognitive storage and processing was less affected. This would be consistent with Kahneman's (1973) conclusion that arousal focuses attention more closely upon a narrower range of cues.

Secondly, they suggest that the effects of the Utility manipulation in Experiment One were due to the arousal elicited by that manipulation. Whether arousal per se enhances the accuracy of behavior perception, it should be noted, should depend upon the discriminability of the relevant behavioral information. If the critical information is easily discriminable, arousal could enhance observer accuracy by restricting interpretation to the few most relevant cues, causing irrelevant information to be screened out. If the critical discriminations for veridical perception are complex, however, arousal could be highly disruptive.

Phase II: Validation of the Units as Perceptual Information

Evidence from the first two studies support the interpretation of behavior segmentation as tapping the preliminary, perceptual base of information in ongoing observation. Consequently, research in this phase of the project focused directly upon establishing the perceptual nature of behavior units. In an earlier study of the reliability of the unit marking procedure, segmentation of a seven-minute sequence was found to be highly reliable over a five-week test-retest interval. If unit formation is indeed a perceptual phenomenon, such reliability would be expected, as segmentation is keyed to specific features of the stimulus field, and it is unlikely that subjects could remember their marking patterns over a five-week interval.

As a first step in this phase, it was decided to replicate that study, to verify that the reliability obtained in that study was not uniquely characteristic of the specific sequence employed.

Experiment Three

Eight different three-minute sequences were constructed, to assess differences in reliability over several different types of behavior. In addition to Fine-Unit and Large-Unit instructional conditions, a Natural Unitization condition was included. Test-retest interval was five weeks.

Method

Subjects

Subjects were twenty-nine males and twenty-eight females recruited from introductory psychology classes at the University of Virginia, and were either paid or given course credit for participation.

Stimuli

Eight sequences were prepared. These consisted of: I. A man pacing impatiently and intermittently answering a phone; II. A man systematically removing stacks of magazines from a table and shelving them; III. A woman performing an interpretive dance; IV. A woman setting a table with plates and food; V. A man clearing a table littered with plates and cups, by knocking them off onto the floor; VI. A man systematically building a tower from tinker toys; VII. A man cheating on a test; VIII. A woman making a series of identical tinker toy constructions and placing them in a pattern on the floor. In constructing these sequences, we attempted to generate sequences that were reasonably diverse, in that they included actions both novel (e.g., VIII) and familiar (e.g., IV), structured (e.g., VI) and unstructured (e.g. III), serious (e.g., VII) and whimsical (e.g., V), planned (e.g., II) and unplanned (e.g., I). Sequences were videotaped; no sound was included. Length of the eight sequences in seconds were, respectively, I: 166; II: 157; III: 151; IV: 156; V: 94; VI: 154; VII: 152; VIII: 198.

Apparatus

Videotapes were presented on an Electrohome 23-inch high resolution monitor, placed four feet from the subjects. Subjects were provided with buttons, as in previous experiments; unit judgements were recorded with an Automated Data Systems 1800E laboratory computer. This permitted precise recording of the timing of unit judgements, and efficient coding of unit data.

Procedure

Procedure was identical to that of the previous study. Condition instructions were identical for Fine-Unit and Gross-Unit conditions. In the additional, Natural-Unit condition, subjects were told, "What I want you to do is to mark off the behavior of the persons you'll be seeing into whatever units seem natural and meaningful to you." All individual differences measures (see below) were taken after the second session. Sequences were presented in a constant order across all conditions and both test sessions.

Design and Analysis

Units were scored in two-second intervals for these data. Selection of interval size for scoring is somewhat arbitrary; Newtson (1973) employed 2.5-second intervals. The criterion for selection in this instance was that less than five percent of all cases yielded multiple marks with this size interval. Data for each subject consisted of the number of units for each sequence and the specific pattern of unit-marking for each subject for each sequence. In addition, subjects completed the I-E scale, Snyder's (1974) self-monitoring scale, and the Marlow-Crowne social desirability scale.

The three indices of reliability were computed, separately for each of the eight sequences. These were: 1) Subject reliability, consisting of the correlations between number of units marked for each sequence at test with number of units marked for the same sequence at retest; these were computed separately for each of the three conditions; 2) Interval reliability, consisting of the correlation between number of marks for a given interval at test and retest, again computed separately by sequence; and 3) Subject X Interval reliability. This index was computed by comparing each subject's two markings of the sequence. The result was, for each subject, a 2 X 2 table giving that particular subject's frequency of matching and mis-matching markings. That is, if an interval was marked both at time one and time two, that was counted as a "correct hit;" not marking the same interval was counted as a "correct miss;" marking at time one, but not time two, was counted as a "false negative;" while marking at time two an interval not marked at time one was counted as a "false positive." Expected frequencies for each of the four cells was computed from the marginals (cf. Siegel, 1956), and subtracted from the observed cell frequencies. The resulting scores for "correct hits" and "correct misses" were then summed, yielding an estimate of the number of matches in the markings exceeding chance for each subject.

A 3 X 2 X 2 X 8 mixed analysis of variance was conducted on total units and on the Subject X Interval reliability index. Factors were Condition (Fine-Unit, Gross-Unit, and Natural-Unit), Sex of Subject, Test (Test vs. Retest), and Sequence (I through VIII).

Subject and Interval reliabilities were averaged within conditions by means of Fisher's r to z transformation and tested for significance (McNemar, 1969). Tests of significance between and within sequences within each condition were computed for both Subject and Interval reliabilities.

Two additional correlation analyses were performed. First, number of units marked for each of the eight sequences were intercorrelated, to test for subject stability in relative unitization rate across the different sequences. Second, number of units for the first 47 intervals of each of the eight sequences were intercorrelated, to test for the possibility that marking patterns consist of some regular pattern regardless of behavior content. Forty-seven intervals were used because the shortest sequence (Sequence V) contains this number of intervals. These correlations may be interpreted as a kind of "alternate forms" reliability index. Due to the volume of data being reported, these matrices are not included; instead, they were averaged via Fisher's z, and those averages reported. As these are interdependent correlations, tests of significance are not appropriate here, so the percent of the correlations exceeding significance at the .05 level are reported.

Finally, the three individual difference measures were correlated with number of units.

Results

Analysis of variance of number of units is reported in Table One. Means by Condition and Sequence are presented in Table Two. Significant effects were observed for Condition ($p < .005$) and Sequence ($p < .005$) and for the Condition by Sequence interaction ($p < .01$). Means tests (Table Two) indicate that the interaction

---

Insert Tables One and Two about here

---

was due to non-significant differences in number of units between the Gross-Unit and Natural-Unit conditions for Sequences I and II. The main effect for Test approached significance ($p < .10$), and was due to a uniform tendency to mark more units at the second marking.

TABLE ONE

Experiment Three

Analyses of Variance of Total Units and Unit Type

Source	df	Total Units MS	Total Units F	Person Chosen MS	Person Chosen F	Situation Produced MS	Situation Produced F
Condition (A)	2	18445.89	6.27***	10512.57	5.36	1323.68	2.59
Sex (B)	1	1123.39	.38	69.65	.04	747.64	1.46
AB	2	1048.63	.36	147.45	.08	579.04	1.13
S (AB)	51	2942.63		1961.35		510.80	
Test (C)	1	1001.78	2.90	621.43	2.39	29.93	.50
AC	2	267.46	.78	336.68	1.29	6.88	.11
BC	1	61.61	.18	.38	.00	41.29	.69
ABC	2	80.78	.23	64.11	.25	54.59	.91
S (AB) C	51	345.06		260		60.26	
Sequence (D)	7	356.85	4.09***	706.92	8.15***	237.18	6.31**
AD	14	190.30	2.18**	260.09	3.00**	46.25	1.23
BD	7	54.84	.63	94.13	1.09	66.76	1.77
ABD	14	40.74	.47	45.60	.53	35.27	.94
S (AB) D	357	87.33		86.71		37.61	
CD	7	68.57	1.81	57.27	1.39	15.66	.81
ACD	14	52.89	1.40	57.31	1.39	10.05	.52
BCD	7	14.52	.38	42.53	1.03	45.15	2.33
ABCD	14	37.41	.99	42.09	1.02	15.03	.77
S (AB) CD	357	37.86		41.32		19.42	

** $p < .01$

*** $p < .005$

TABLE TWO

Experiment Three

Total Units by Sequence and Instructional Condition

	Sequence							
Condition	I	II	III	IV	V	VI	VII	VIII
Fine Unit	18.30^{a}	16.78^{a}	27.63^{a}	21.34^{a}	20.95^{a}	23.45^{a}	20.00^{a}	26.23^{a}
Gross Unit	7.47^{b}	3.77^{b}	4.00^{b}	6.32^{b}	6.03^{b}	5.94^{b}	6.03^{b}	8.27^{b}
Natural Unit	11.90^{b}	8.73^{b}	12.25^{c}	14.90^{c}	12.05^{c}	11.00^{c}	13.58^{c}	12.25^{b}

Note: Means within Sequence and Unit measure with different subscript differ, $p < .05$.

$Tot_{LSD} = 5.94$

The three reliability indices are reported in Table Three.

Insert Table Three about here

Average Subject reliability by condition was .76, .63, and .85 for the Fine-Unit, Gross-Unit, and Natural-Unit conditions, respectively, and all were significant ($p < .01$). Tests of significance between correlations within and between sequences failed to yield any significant differences in Subject reliability.

Average Interval reliabilities by condition were .61, .62, and .63 for Fine-Unit, Gross-Unit, and Natural-Unit conditions, respectively ($p < .01$). Two sequences yielded significant differences between conditions in Interval reliability. Interval reliability for the Natural-Unit condition in Sequence III, the dance sequence, was significantly lower than Interval reliabilities for the Gross-Unit and Fine-Unit conditions ($p < .05$), which did not differ from each other. In Sequence VIII, in which a woman moved about constructing a series of figures, Natural-Unit reliability was significantly greater than Fine-Unit reliability ($p < .05$); Gross-Unit reliability fell between the two and was not reliably different from either of the other two conditions.

Within conditions, many differences in Interval reliabilities for the eight sequences were observed. Within the Fine-Unit condition, Sequence IV, the table-setting sequence, was significantly more reliable than all but Sequence I, the telephone-answering sequence ($p < .05$). Sequence I was not significantly different from the others. Within the Gross-Unit condition, Sequences I and IV were significantly more reliable than the others; in addition, Sequence VIII was significantly more reliable than Sequences III and VI (depicting the construction of a block tower) ($p < .05$). Sequences differed strongly in the Natural-Unit condition with both IV and VIII significantly more reliable than II, III, and VI; interval reliability was greater for IV than V, and V, VI, and VII were more reliable than III. It should be noted that degrees of freedom for Subject reliabilities are a function of number of subjects, while degrees of freedom for Interval reliabilities depend upon the number of intervals. The analysis of Interval reliabilities is thus more powerful.

Means for the Subject X Interval reliability index, indicating the number of precise matches exceeding chance over the five-week interval, are also reported in Table Three. Analysis of variance indicated significant main effects for Condition ($F = 5.81$, $df = 1/51$, $p < .05$), Sequence ($f = 7.07$, $df = 7/357$, $p < .001$) and a Condition by Sequence interaction ($f = 2.07$, $df = 7/357$, $p < .01$).

TABLE THREE

Experiment Three

Three Indices of Reliability of Unit Marking for Eight Sequences at Three Levels of Unit Instruction

Condition	Reliability Index	Sequence 1	2	3	4	5	6	7	8
Fine Unit	Subject	.75***	.60	.74***	.75***	.69***	.91***	.78***	.69***
	Interval	.68***	.51***	.59***	.77***	.63***	.47***	.60***	.57***
	Subject X Interval	5.12*	3.13*	2.71*	4.80*	3.23*	3.74*	4.36*	10.46*
Gross Unit	Subject	.61**	.36	.58	.79***	.72***	.52	.73***	.60**
	Interval	.80***	.61***	.42***	.80***	.48***	.37***	.60***	.66***
	Subject X Interval	3.05*	1.51	1.69*	2.33*	2.78*	1.81*	1.81*	2.31*
Natural Unit	Subject	.85***	.93***	.63	.92***	.78*	.94***	.89***	.61
	Interval	.71***	.40***	.13	.80***	.61***	.57***	.75***	.77***
	Subject X Interval	2.60*	.77	.44	3.10*	2.29*	2.21*	4.68*	6.36*

*p < .05

**p < .01

***p < .001

Note: H_o for all significance tests is that r of $\bar{x} = 0$.

Subsequent means test indicated that Subject X Interval reliabilities for all eight sequences were significantly greater for the Fine-Unit condition than for the other two conditions; this index differed between the Gross-Unit and Natural-Unit conditions only for Sequence VIII, where significantly more matches were observed in the Natural-Unit condition.

More importantly, when compared to zero, only three of the twenty-four means failed to reach significance (Fisher's LSD = 1.60, df = 357, $p < .05$, one-tailed).

Intercorrelations between number of units marked for each of eight sequences averaged .77 for the Fine-Unit condition (96 percent were significant), .69 for the Gross-Unit condition (68 percent were significant), and .79 for the Natural-Unit condition (98 percent were significant). This pattern of results implies a substantial contribution of a "characteristic rate" of unitization across a diverse set of behavior sequences.

Interval intercorrelations for the first 47 intervals were also computed between sequences, to rule out contributions to reliability of regular marking patterns independent of behavior content. These intercorrelations averaged .16 in the Fine-Unit condition (13 percent were significant), .04 in the Gross-Unit condition (5 percent were significant), and .03 in the Natural-Unit condition (11 percent were significant). Overall, about 10 percent of these correlations were significant; the range was from .45 to -.32 with the majority close to zero.

Correlational analysis of the relation between unitization and the three individual differences measures failed to yield any meaningful or consistent patterns of results.

Discussion

Results of this second study of marking reliability are consistent with those of the first, and demonstrate that unitization for both subjects and behavior sequences are substantially reliable across a range of behavior sequences. In addition, comparison of natural-unit results to those of unitization extremes indicate comparable reliability, suggesting that these instructional variations may not be extremely atypical in the perceptual experience of individuals. This result, in turn, is consistent with the notion that individuals have a "range of analysis" in behavior perception, and that level of analysis selected at a given time may be dependent on situational factors.

The fact of such reliability, it should be noted, does not bear on the assumption that perceptual organization of action is relatively constant across observers. It does demonstrate that the measure is a reliable and substantially precise one in that, despite variability across observers, the measure displays considerable stability in representing the operation of the same observer at different times.

With respect to our goal in this phase of the project of establishing that behavior units are perceptual units, this pattern of stability within variability is consistent, as least, with that interpretation. If behavior units are perceptual, then they are formed on the basis of information available in the stimulus at the time of their formation. That behavior units may vary across observers, or due to situational factors or instruction, is not inconsistent with this notion. Implicit in the present approach is the assumption that behavior perception is an active, selective, perceptual process. If variations in observer skill are due to variation in the perceptual basis of their judgements, at least two conditions must hold. First, segmentations must have a basis in the stimulus, as noted. This is simply to say that the perceptual organization of a given sequence of behavior is not arbitrary with respect to the stimulus. That perceptual organization is constrained by the stimulus, however, does not mean that it is wholly determined by the stimulus, and hence constant across observers. Rather, and this is the second condition for there to be a relation between observer skill and behavior segmentation, it must be possible for segmentation to vary within limits set by the stimulus. This issue will be taken up in more detail at a later point. It is raised here only because, in addressing the first question, we shall deal with normative segmentations of sequences with relatively unambiguous unitary interpretations. Evidence for consistency across observers under these conditions, therefore, should not be taken as evidence ruling out observer differences in perceptual organization of behavior.

Fodor and Bever (1965) note that there are a number of techniques for experimentally determining the perceptual organization of a complex stimulus. The simplest method is to appeal directly to the intuitions of the perceiver. This method is essentially the same as the unit-marking technique we have been using. Verification that the technique is indeed tapping perceptual organization, however, requires convergent evidence from an alternative technique. In this connection, Fodor and Bever (1965) note that, "A more subtle way of establishing the segments of a complex percept exploits the tendency of a perceptual unit to preserve its integrity by resisting interruptions (p. 326)." The next study adopted this second strategy, demonstrating that the units identified by the unit marking procedure are indeed more resistant to disruption

within that action unit than at the boundary points between actions.

Experiment Four

A series of brief filmed episodes were prepared, and consensus points of division and continuity were identified. We shall term those intervals or points in the sequence that are most likely to be marked as segmentation points "breakpoints," as they are the points at which the behavior stream is broken up into its parts. Intervening intervals or points we shall term "nonbreakpoints." Sections of film were then cut out of these intervals and the film was spliced back together. These films were then presented to subjects, whose task was to detect any and all occurrences of missing action. If the units identified by the unit marking procedure are true perceptual units, then they should resist interruption. Therefore, there should be poorer detection of missing frames at nonbreakpoints (within the perceptual unit) than at breakpoints (at unit boundaries). In addition, evidence that the units identified through the use of the unit-marking technique are similar to those employed by an alternate non-marking group would indicate that the technique does not interfere with the process it is designed to measure.

Method

Subjects

Subjects were twenty-one persons (ten males, eleven females) who were recruited in the Charlottesville area and paid for participation in the experiment.

Stimuli and Item Selection

Nine 30-second action sequences were prepared for use in the experiment, and segmented by twenty subjects according to the Newtson unit-marking procedure.

The nine sequences were recorded on 16 mm. black and white film. These consisted of: 1) A man nervously leafing through a magazine; 2) A man working on a radio, and smashing it in frustration; 3) A woman cutting out a dress pattern; 4) A man repairing a motorcycle; 5) A woman accidentally spilling a cup of coffee; 6) A man searching for a lost item in a desk; 7) A man setting out tools; 8) A man pacing and then rushing to answer a phone; and 9) A man cheating on an exam.

Breakpoint and nonbreakpoint intervals were then identified from the unit markings provided by the pre-test group. Unit marks

were tabulated for each one-second interval of each sequence, and three breakpoints and three nonbreakpoints were identified in each sequence. To select breakpoint and nonbreakpoint intervals from the pretest group segmentations the total number of units recorded for each sequence by all subjects was first divided by the number of intervals, yielding a mean number of marks per interval. The standard deviation of marks per interval was then calculated for each sequence. Three intervals with total marks at least one standard deviation above the mean were selected as breakpoints, and three intervals with total marks more that one standard deviation below the mean were selected as nonbreakpoints. An additional constraint on interval identification was that, from each sequence, breakpoints (BP) and nonbreakpoints (NBP) alternate. That is, actual order was BP-NBP-BP-NBP-BP-NBP for the intervals from four action sequences and NBP-BP-NBP-BP-NBP-BP for the remaining five sequences.

Four frames were deleted from one of the three breakpoint intervals and from one of the three nonbreakpoint intervals; eight frames were deleted from a second interval of each type; and twelve frames were deleted from the third interval. The particular intervals from which these sections were removed were determined randomly for each sequence. As projection speed was 24 frames per second, duration of deletions was 1/6, 1/3, or 1/2 second. The nine sequences were then spliced together with a five-second blank between each sequence.

Apparatus

Responses were recorded by means of a button mounted on a 5 x 8 x 2 inch box, connected to an Automated Data Systems 1800E laboratory computer. The times at which judgements were made were recorded by the computer for later scoring.

Procedure

Subjects were run in pairs. They were seated at a table facing the projection screen, separated by a partition. To preclude their influencing each other's responses, they wore earphones, and response boxes were cushioned with foam rubber.

Subjects were instructed as follows: "In this experiment, I am going to show you nine short films of people doing a variety of things. What we are interested in here is how essential various parts of human action are for the perception of continuous behavior. For this reason, we have eliminated certain parts of the action from the films you are about to see. We did this by simply cutting out

either a small or large number of frames at various points in the films."

"What I'd like you to do is to watch the films closely and to mark the points where you notice some action or action part missing from the film. To do this, simply press the red button before you."

Results

Data from each subject consisted of a series of times at which deletions were detected. These were compared to the actual times at which deletions occurred, and counts were made of the number of each type detected. Criterion for accurate identification was that the subject indicated a deletion within one second after the actual time of deletion. This resulted in a total of 23 inaccurate responses, or an average of 1.10 false identifications per subject.

The mean number of deletions detected by subjects is shown in Table Four as a function of Interval Type (Breakpoint vs. Nonbreakpoint) and Number of Frames Deleted (Four, Eight, or Twelve).

Insert Table Four about here

Analysis of these data yielded a significant main effect of Interval Type ($F(1,20) = 117.92$, $p < .001$), and Number of Frames Deleted ($F(2,10) = 19.24$, $p < .001$), and a significant interaction of these factors ($F(2,40) = 19.82$, $p < .001$). Subsequent t-tests indicated that detection improved significantly as a function of size of deletion for breakpoints only.

The hypotheses were thus confirmed, in that deletions at breakpoints were detected significantly better than deletions at nonbreakpoints at all sizes of deletion.

Discussion

Results clearly support the view that ongoing behavior is perceived in units, and that the unit marking technique can be used to measure the perceptual unit of ongoing behavior. It might be argued, however, that these results may be artifactual because of differences in the stimulus at the two types of intervals. For example, it could be that breakpoint intervals occur during periods of considerable movement, while nonbreakpoint intervals occur during periods of relative immobility. Deletions during breakpoint intervals would thus be easier to detect, and those during nonbreakpoint

Table Four

Experiment Four

Number of Deletions Detected by Interval

Type and Number of Frames Deleted

Interval Type	Number of Frames Deleted		
	4	8	12
Breakpoint	3.95^{b}	5.48^{c}	7.00^{d}
Nonbreakpoint	3.24^{a}	2.91^{a}	3.38^{a}

Note: Mean is number deletions detected out of 9 possible.

Means with different superscripts differ, $p < .001$.

intervals relatively more difficult. As these intervals were identified by subjects with the unit-marking procedure, subjects could have been responding to this "superficial" aspect of the stimulus, and not recording true perceptual units.

The stimulus tapes in the present experiment, however, portrayed nearly constant movement, and thus a simple movement-no movement interpretation is unlikely. The possibility remains that more subtle differences between the two interval types do exist. The fact of physical differences between the interval types is a problem only if those differences (1) are not the actual basis for perceptual organization, and (2) do indeed render deletions more or less detectable. That is, the perceptual meaning of ongoing behavior must have some objective basis in the stimulus; unit boundaries should thus be expected to differ from intervals within the unit on some physical dimension. The nature of the interaction between the stimulus and the perceptual process in forming perceptual units is secondary to the question of whether such behavior units exist. What is of concern, then, is whether these data fully establish that existence. If it can be demonstrated that the interval types have different properties with respect to the comprehension of meaning in the behavior, then the present data can be accepted as establishing the reality of behavior units.

Experiment Five

Two alternate interpretations of the results of the preceding study are plausible. One would hold that the units of behavior perception are formed in the intervals between the unit boundaries, and perceivers are thus less sensitive to disruption in these intervals. Inspection of breakpoints themselves, however, suggests an alternative view. A series of breakpoints conveys an almost comic strip quality, in that they appear to summarize an event very well. Nonbreakpoints, on the other hand, appear highly ambiguous, in that a large number of alternative constructions of the event appear to be consistent with them. It is possible, therefore, that greater sensitivity to deletions at breakpoints occurred because the deletions interfered with unit formation. If unit formation occurs at breakpoints, then a series of breakpoints extracted from a film and viewed in succession should provide a more adequate and understandable summary of the action sequence than a comparable series of nonbreakpoints.

In the present study, subjects viewed eight series of three successive breakpoints or three nonbreakpoints (selected from the intervals between breakpoints), and then described the action portrayed and the degree to which the sequence was intelligible, or

comprehensible. Criterion for accuracy of description was the extent to which descriptions matched those of a control group who viewed the continuous behavior. If comprehension of ongoing behavior is organized at breakpoints, then subjects viewing breakpoints should describe the action more accurately than subjects viewing nonbreakpoints. In addition, subjects viewing nonbreakpoints should rate the portrayed action as less intelligible than subjects viewing either breakpoints or the continuous sequences.

Order of presentation of the stimuli was also varied. If breakpoints are the basis for perceptual organization of a sequence of action, they should also contain information as to the order in which the event occurred. Accordingly, stimuli were presented in correct or incorrect order, and subjects were asked to judge whether each slide set was presented in correct order. These judgements were expected to be more accurate when the slides portrayed behavior at breakpoints.

Method

Subjects

Subjects were seventy-nine undergraduates from introductory psychology classes at the University of Virginia (thirty-four males, forty-five females). Subjects received either course credit or payment for participation.

Stimuli

Stimuli were eight of the nine 30-second action sequences employed in the previous experiment. The cheating sequence was not employed in this study.

Breakpoint and nonbreakpoint intervals were the same ones used in the previous experiment. Three breakpoints and three nonbreakpoints were selected from each sequence, yielding 48 items, 24 of each type. A single frame from the center of the interval was extracted and mounted as a slide.

Apparatus

Slides were presented with a Kodak Carousal slide projector programmed to advance every five seconds, at a distance of approximately eight feet. Projected size was 24 inches on the diagonal.

Measures

Subjects responded to the question, "How intelligible (understandable) were the three slides taken as a whole. That is, do the slides depict an intelligible event?" by means of a nine-point scale ranging from "not at all intelligible" to "very intelligible." Subjects were then asked to give a one-sentence description of the behavior portrayed. These two measures were also administered to the pretest group who saw the continuous sequences, with appropriate modifications.

Subjects were also requested to indicate whether or not the slides were presented in the correct order, and, if not, to give the correct ordering.

Procedure

Subjects were run in groups of nine to twelve persons. They were informed that they would view eight sets of three slides, and were instructed on how to complete the set of measures for each of the eight slide sets. Each slide was presented for five seconds. After presentation of the three consecutive slides in a set, subjects completed the measure; when all were ready, the next set was shown. Upon completion of the slide sets, the purpose of the study was explained and subjects were dismissed.

Design and Analysis

Data for each subject consisted of intelligibility ratings, descriptions, and order judgements.

Descriptions were scored for accuracy according to protocols developed from the descriptions of the pretest group. Two independent raters blind as to condition scored each description for correct inclusion of features of the action (0-3) and for correct order of the action (0-1). These ratings were performed with 94.5 percent agreement. One rater's scores were thus selected as data for the analysis. The two ratings were summed, yielding an accuracy index ranging from zero to four.

Design was a 2 X 2 X 2 repeated measures analysis of variance. Variables were Order of Presentation (Correct vs. Incorrect), Slide Type (Breakpoint vs. Nonbreakpoint), and Item Set, with two of the sequences in each condition. Four independent groups were run, so that all slide sets appeared in each condition. That is, while subjects in one group saw Sequences One and Two in the Breakpoint-Correct Order condition, a second group saw Sequences One

and Two in the Breakpoint-Incorrect Order condition, and so on. The result was that while each subject saw only one of the four slide sets from a given sequence, each sequence appeared in all conditions in the design. Data were then combined across groups for the analysis of variance. This design was employed for analysis of accuracy, intelligibility ratings, and order judgements. Re-ordering data were analyzed in a similar design, dropping the Order of Presentation factor.

Results and Discussion

Analysis of variance of intelligibility ratings yielded a main effect of Slide Type ($F(1,75) = 53.39$, $p < .001$) and a Slide Type X Order of Presentation interaction ($F(1,78) = 20.86$, $p < .001$). Means are presented in Table Five. Subsequent t-tests indicated

---

Insert Table Five about here

---

that intelligibility differed as a function of Presentation Order for Breakpoints only. The hypothesis that breakpoints are more intelligible than nonbreakpoints is thus confirmed. The pretest control, who viewed the continuous behavior sequences, gave a mean intelligibility rating for the eight sequences of 7.25. This was not significantly different from the mean of the Breakpoint-Correct Order condition ($t(237) = .26$); the other three conditions were rated as significantly less intelligible in comparison to this control (t's$(237) = 3.96$, 5.00, and 5.45, $p < .05$), for the Breakpoint-Incorrect Order, Nonbreakpoint-Correct Order, and Nonbreakpoint-Incorrect Order conditions, respectively.

Analysis of variance of description accuracy yielded significant main effects for Slide Type ($F(1,78) = 63.91$, $p < .001$). Order of Presentation ($F(1,78) = 6.33$, $p < .05$), and Item Set ($F(1,78) = 5.22$, $p < .05$). Means are reported in Table Five. Breakpoints were more accurately described than nonbreakpoints, thus confirming that hypothesis. Correctly ordered slides were also more accurately described than incorrectly ordered slides. The interaction between Slide Type and Presentation Order ($F(1,78) = 1.67$) did not attain significance, as it did on the intelligibility ratings.

Means for order judgements are presented in Table Five as well. Analysis of variance indicated a significant main effect of Slide Type ($F(1,78) = 95.31$, $p < .001$) and a Slide Type by Order of Presentation interaction ($F(1,78) = 5.66$, $p < .05$).

TABLE FIVE

Experiment Five

Results on Three Measures by Order of Presentation and Slide Type

Measure	Slide Type			
	Breakpoints		Nonbreakpoints	
	Order of Presentation		Order of Presentation	
	Correct	Incorrect	Correct	Incorrect
Intelligibility	7.32^{a}	6.18^{b}	4.95^{c}	5.45^{c}
Description Accuracy	2.29^{a}	2.00^{a}	1.58^{b}	1.49^{b}
Order Judgements	$.83^{a}$	$.77^{a}$	$.37^{b}$	$.47^{c}$

Note: Means within each measure with different superscripts differ by t-test, $p < .05$.

Finally, when Order was accurately judged to be incorrect, breakpoints were correctly re-ordered 46 percent of the time, while nonbreakpoints were correctly re-ordered with only 14 percent accuracy ($F(1,78) = 40.82$, $p < .001$).

Results confirmed that breakpoints are the basis for the formation of perceptual units of behavior. Clearly, the units identified by the unit marking procedure are not selected according to an arbitrary criterion, but are significantly related to the meaning of the behavior. These results indicate that it may be more accurate to view ongoing behavior as consisting of transitions between successive points of definition, with meaning dependent upon the nature of the transitions, than in terms of bounded segments, with meaning a function of the content within those boundaries (cf. Barker, 1963). At minimum, these data indicate that some intervals of ongoing behavior are more distinctive for observers than others.

Experiment Six

Given their distinctiveness, it follows that recognition memory for breakpoints should be superior to recognition for nonbreakpoints. Evidence that this is the case would add converging evidence for the psychological reality of perceptual units of behavior.

In the present study, a series of films were prepared, divided in half, and marked by a pretest group. Breakpoints and nonbreakpoints were identified, extracted, and mounted as slides. Subjects viewed one half of each film, and then judged whether or not slides drawn from both film halves came from the film they had just seen. The design was counterbalanced so that each item was an "old" item for half of the subjects, and a "new" item for the other half.

In addition, conditions of viewing were varied. In one condition subjects segmented the film while viewing; in another, subjects simply watched the film. This was done to insure that any recognition differences found were not artifacts of the marking response provided only at breakpoints. It was predicted that recognition memory for breakpoints is superior to recognition memory for nonbreakpoints.

Method

Stimuli

Six two to six-minute behavior sequences were filmed. Sequences consisted of: 1) A man taking a test; 2) A woman dancing;

3) A woman cutting out a dress pattern; 4) A man repairing a motorcycle; 5) A man looking for a lost object; and 6) A man waiting impatiently for a phone call.

These films were cut in half, and segmented by twenty standardization subjects instructed to record units of behavior that seemed "natural and meaningful to them." To check on the possibility that unitization would differ for one half of a sequence as a result of prior viewing of the other half, ten subjects segmented the first halves of all six sequences and then the second halves, while the order was reversed for the other ten standardization subjects. No differences in marking pattern was detected. Breakpoint items were selected independently from each sequence half. Markings were tabulated for one-second intervals, and break and nonbreakpoint were identified as in Experiment One. Three break and nonbreakpoints were selected randomly from those meeting the criterion for each film half. A single frame from the center of these intervals (24 frames per second) was extracted and mounted as a slide in special 16 mm. slide mounts. This yielded 36 breakpoint and 36 nonbreakpoint slide recognition items.

Subjects

Subjects were sixty-four undergraduates enrolled at the University of Virginia (twenty-six males, thirty-eight females), who received either payment or course credit for participation.

Apparatus

Films were projected on a 4 X 5 screen at a distance of eight feet from the subjects. Slides were shown via a Kodak Carousel slide projector programmed to advance every 15 seconds. For those subjects who marked the sequence (see below), unit data were recorded by a push-button connected to an Automated Data Systems 1800E laboratory computer programmed to record the time of each signal.

Procedure

Subjects in the Marking condition were seated at a table in pairs (separated by a partition), and informed that they would see six short films. They were further advised to be prepared to answer some questions about them at a later time. Following standard instructions on the marking procedure (Newtson, 1973), subjects were told, "What I would like you to do is mark off for me the naturally-occurring meaningful actions in the sequences, as you see them. To do this, simply press the red button before you.

Let me emphasize that there are no right or wrong ways to do this; I just want to know how you do it."

Following presentation of the film sequences, subjects worked on a filler task (completion of Snyder's 1974 Personal Reaction Inventory) for ten minutes. The recognition test was then administered. Subjects were told that they would be shown slides, some of which were extracted from the sequences they had seen, and some of which were extracted from sequences employing the same actor and situation but which they had not seen. A sample of old and new items were then projected for 15 seconds each and subjects recorded yes-no judgements and gave confidence ratings on a 0 - 3 scale. The procedure was the same for the remaining 72 items.

Procedure for the Watch condition was the same. Subjects were run in groups of three to eight. Instead of marking instructions, these subjects were told, "What I want you to do is to watch the people in the sequence very carefully. Try not to miss any of the meaningful actions they perform." Following the viewing of the films subjects in this condition performed the same ten-minute filler task and took the recognition task previously described.

Half of the subjects viewed the first halves of the six sequences, and the other half viewed the second halves. All subjects then judged all 72 items, presented in two random orders, with the constraint that no items from the same sequence were presented successively. Thus 36 items (18 breakpoints and 18 nonbreakpoints) were "old" items, and 36 were "new" for each group.

Design and Analysis

Data for each subject consisted of 72 yes-no judgements and corresponding confidence ratings for those judgements. These were scored for accuracy, and proportion of correct judgements were obtained for each subject for old and new breakpoint and nonbreakpoint items, yielding four scores for each subject. Accuracy scores were analyzed in a 2 X 2 X 2 X 2 mixed analysis of variance design. Between-subjects factors were Conditions of Viewing (Mark vs. Watch) and Film Half (First vs. Second). Repeated measures were Prior Exposure (Old vs. New) and Slide Type (Breakpoint vs. Nonbreakpoint).

In addition, a signal detection analysis was performed to verify the greater discriminability of breakpoint items. One standard procedure incorporates both yes-no judgements and confidence ratings to compute a d' for each subject, and these values are then analyzed by analysis of variance. This was not possible for the present data, however, because some combinations of yes-no judgements and confidence ratings did not occur for some subjects.

Accordingly, a method described by Snodgrass (1975) for computing an alternative, nonparametric index of sensitivity, A' was employed. This method computes the sensitivity level for each subject from the signal and noise distributions alone. The A' index has been found to be comparable to more conventionally computed d' values (Snodgrass, Volvovitz, & Walfish, 1972). These values were analyzed with a two-way analysis of variance comparing Breakpoints to Nonbreakpoints.

Overall d' values were computed (i.e., over the entire sample, not subject by subject), and are reported. In comparing A' and d' values, one should keep in mind that A' is a percent index, while d' is a z-score index of discriminability.

Results

Analysis of variance yielded significant main effects for Film Half ($F = 30.11$, $df = 1/60$, $p < .005$), Prior Exposure ($F = 10.33$, $df = 1/60$, $p < .005$), and a significant three-way interaction of these variables ($F = 4.47$, $df = 1/60$, $p < .05$). No effects involving Conditions of Viewing were significant ($p < .10$).

The main effect on Slide Type confirmed the hypothesis, in that Breakpoints were significantly better recognized than Nonbreakpoints. Confirmation was not clear-cut, however, as tests of the three-way interaction (see Table Six) indicated that recognition

---

Insert Table Six about here

---

accuracy for Breakpoints and Nonbreakpoints did not differ for New items from the first half of the films. The predicted difference was obtained at the other three levels of Film Half and Prior Exposure, however.

The overall difference between the film halves was probably due to content difference in the films. For some reason, more units were identified in Second halves of the films than in First halves (10.20 vs. 8.70, $t(1226) = 2.50$, $p < .01$).

Results from the signal detection analysis confirmed that higher recognition for Breakpoints is not due to a positive response bias and reflects real differences in discriminability. Analysis of variance of individual A' values yielded a significant difference between Breakpoints ($\bar{x} = .76$) and Nonbreakpoints ($\bar{x} = .70$; $F = 18.00$, $df = 1/63$, $p < .001$). The overall d' values were .96 for Breakpoints and .59 for Nonbreakpoints.

Table Six

Experiment Six

Recognition Accuracy by Prior Exposure,

Slide Type and Film Half

Film Half	Prior Exposure			
	Old Slide Type		New Slide Type	
	Breakpoint	Nonbreakpoint	Breakpoint	Nonbreakpoi
First Half	.695[a]	.610[b]	.570[c]	.568[c]
Second Half	.761[d]	.729[c]	.674[f]	.636[g]

Note: Means with different superscripts differ by *t*-test, $p < .05$.

Despite the small absolute size of the differences obtained, they were highly consistent. Tabulation of the number of subjects yielding the predicted pattern of results, superior recognition for Breakpoints than Nonbreakpoints, indicated that data from 62.5 percent of the subjects were in accord with predictions. Twenty-five percent showed no difference in accuracy as a function Item Type, and only 12.5 percent reversed the prediction. A sign test indicated that this difference was highly significant ($p < .001$).

It must be remembered that the present study employed breakpoints most consistently identified by a group. Presumably, if the individual breakpoints marked by each subject had been employed as recognition items, results would have been more powerful. Evidence from the study of the reliability of unit marking (Experiment Three) indicates quite clearly that the bulk of between-subjects variability in marking patterns is due to stable individual differences in unitization, rather than measurement error; item selection was thus less than perfect here. Internal analyses of the present data are wholly consistent with this notion. Since all reversals in predicted pattern of results and five of the eleven "no difference" results occurred in the Mark condition (suggesting some effect of Conditions of Viewing), it was possible to compare their unitization with that of the pretest group; markings for the subjects in the Mark condition who supported predictions were also compared. Subjects whose recognition data did not support the hypothesis coincided with the pretest group (who provided the basis for item selection) on an average of only 4.8 out of 18 breakpoints; this mean for subjects who supported the hypothesis was 10.8 ($t = 3.5$, $df = 30$, $p < .01$).

Finally, it should be pointed out that the differences in these results were quite consistent across items, and were not due to the impact on the overall means of any small subset of items used in the study.

Results from a pilot test of the recognition procedure showed more dramatic differences in recognition between Break and Nonbreakpoints. Twelve of the pretest subjects, who had marked all of the sequences, were recalled two weeks after viewing the sequences. They were informed that their task was to discriminate old and new items. Proportion of New items to be expected was not specified; in fact, all items were "Old" for this pretest group. These subjects correctly identified 76 percent of Breakpoints as Old items; recognition accuracy for Nonbreakpoints, on the other hand, was only 45 percent, a difference significant at the .001 level. The strength of these results may be due in part to the advantage that this group had in item selection.

Discussion

The results of these last three studies clearly validate the unit marking technique as tapping the perceptual organization of ongoing behavior. The fact that the unit boundaries, or breakpoints, identified by one group convey important information for alternate groups of observers indicates that the technique is relatively free of interference with normal processes of observation, as well.

In addition to validating the perceptual nature of behavior units, furthermore, research in this phase of the project provides an essential starting point for a theoretical model of ongoing observation: that perceptual organization proceeds by the discrimination of successive points of definition in the ongoing stimulus. This finding provides the basis for the next phase of the current investigation.

Phase III: Towards a Theoretical Model of Observational Processes

Research in this phase focused upon the bases of the perceptual organization of ongoing behavior. Evidence from the preceding studies demonstrate that we have succeeded in identifying the subjective unit of action. By specifying the objective stimulus conditions that give rise to these units, then, we may arrive at a point from which the nature of the process may be seen more clearly.

Experiment Seven

As noted previously, the perceptual organization of ongoing behavior must have some objective bases in the stimulus. Specification of the bases of perceptual organization, therefore, requires an adequate understanding of the objective stimulus characteristics of ongoing behavior itself. The findings that breakpoints contained more of the information from the continuous sequence than nonbreakpoints (Experiment Five) imply that behavior, as a stimulus, varies considerably from moment to moment in the amount of information available for its interpretation. That the information value of stimulus intervals plays a role in unit formation is consistent with the findings that perceivers are more sensitive to disruption during these same intervals during ongoing observation (Experiment Four) and that these intervals are more salient in memory (Experiment Six). Whatever the objective basis of the stimulus information is, therefore, it is highly variable within the behavior stream in that more of it exists at breakpoints than at nonbreakpoints.

One possibility is that actions are defined by the achievement of distinctive states by the actor which are, in and of themselves, meaningful. Breakpoints themselves, in this view, would define actions. The most distinctive characteristic of ongoing behavior, however, is change over time. A second, perhaps more likely, interpretation is that actions are defined by the state-to-state changes depicted by successive breakpoints. That is, the distinctiveness of breakpoints would be due to a distinctive change having occurred, rather than a distinctive state having been achieved.

The issue is, in a sense, whether breakpoints are selected according to an absolute or a relative property of ongoing behavior. According to a "distinctive state" hypothesis, stimulus points are marked as breakpoints because they have an absolute property, meaningfulness, independent of previous meaningful stimulus points. According to a "distinctive change" hypothesis, breakpoints in and of themselves would not be distinguishable from other stimulus points; their distinctiveness would be due entirely to their contrast with the point selected as the previous breakpoint.

One means of testing these notions is to compare the positions of the actor in a given sequence at breakpoints and at nonbreakpoints. If breakpoints represent the achievement of distinctive states of the actor, then they should consist of actor positions that are, on average, different from positions at nonbreakpoints. That is, if one randomly compared actor positions between breakpoints and nonbreakpoints, the distinctive state hypothesis should predict a greater average difference in actor position between breakpoints and nonbreakpoints than between position at paired stimulus points chosen at random. To the extent that breakpoints are selected to be nonredundant with other positions of the actor, in other words, they should differ from stimulus points not selected on this basis. If a distinctive change hypothesis is correct, however, no such difference should be observed.

Accordingly, segmentations of seven behavior sequences obtained in a previous study of the reliability of the unit marking procedure (Newtson, Engquist & Bois, 1976) were used to identify breakpoints and nonbreakpoints in those sequences, and position of the actor was coded at those points. Breakpoints were then randomly paired with nonbreakpoints, and codings were compared to obtain an index of difference in position between each pair of points. An equal number of pairs of stimulus points from the same action sequence were also selected, on a random basis, to provide a baseline for the comparison.

If breakpoints are selected on the basis of a distinctive change in position having occurred relative to the previous breakpoint, however, it is possible to make the following predictions from the distinctive change hypothesis. First, the distinctive change hypothesis requires that, when the position of the actor at each breakpoint is compared to position at the next breakpoint, degree of change be greater than degree of change, on average, within action units (i.e. between successive nonbreakpoints). The The one difficulty with this prediction is that unit boundaries are more separated in time than successive points within the unit. It should be noted, however, that separation in time does not necessarily insure a large difference in actor position, as many positions constantly recur. If this prediction is not supported, therefore, this hypothesis is clearly disconfirmed. It is possible, however, to control for this factor by selecting nonbreakpoints for comparison that are matched to the breakpoint pairs for separation in time. This requires inclusion of nonbreakpoint comparisons from different action units, between which a distinctive change has presumably occurred; the hypothesis would still predict a greater degree of change between breakpoint pairs, as nonbreakpoint comparisons may consist of redundant pairs, while breakpoint pairs may not.

In addition, more specific predictions may be made from the meaningful change hypothesis concerning the degree of change within the action unit. If this hypothesis is correct, a point is marked as a breakpoint because a distinctive change relative to the previous breakpoint has occurred. The preceding stimulus point within the action unit, however, was not marked presumably because it did not contain the critical change. One could predict, therefore, a greater degree of change between nonbreakpoints immediately preceding breakpoints and those breakpoints than between successive nonbreakpoint pairs within the action unit.

Similarly, the meaningful change hypothesis would predict that when breakpoints are compared to their immediately following nonbreakpoints, degree of change should be, on average, lower than that between either breakpoint pairs or nonbreakpoint-to-breakpoint pairs.

The main predictions of both hypotheses were tested with the segmentations obtained from subjects instructed to divide the behavior sequences into "whatever actions seem natural and meaningful to you." As findings on variations in level of perceptual analysis (Frey and Newtson, 1973; Newtson, 1973; 1976) indicate, however, there is no one "true" or correct segmentation for a given sequence of behavior, but rather a range of possible organizations. That

perceptual organization is based upon certain changes or states in the ongoing stimulus does not necessarily imply that the perceptual process must be exhaustive with respect to those changes. That is, the presence of a meaningful transformation in the stimulus may be a necessary but not sufficient condition for unit formation. A given behavior sequence may contain the information to support a variety of constructions of its content. The more specific predictions of the meaningful change hypothesis for degree of change within the action unit, therefore, were tested with the segmentations of subjects instructed to divide the behavior sequences into as small, and as large,units of action as seemed natural and meaningful, as well as on segmentations obtained under natural-unit instructions. If the hypothesis is correct, the predictions should be confirmed at these instructional extremes as well.

Predictions of the distinctive state hypothesis for this analysis are not as clear. If breakpoints are selected on the basis of an absolute, rather than relative, property, there is no requirement that successive breakpoint pairs differ among themselves more than nonbreakpoint pairs equidistant in time. If the distinctive state hypothesis is correct, therefore, no such difference should be observed. If, however, breakpoints do consist of distinctive positions, one might predict, on average, that a relatively high degree of change might be observed between nonbreakpoints and immediately following breakpoints, as the actor moves into a distinctive state, as well as between breakpoints and immediately following nonbreakpoints, as the distinctive state is vacated. Change within the action unit, between successive nonbreakpoints, should equal, on average, the average degree of change between any two adjacent stimulus points.

Method

Behavior Sequences

Seven sequences were selected for analysis from the previous study of the reliability of the unit marking procedure (Newtson, Engquist, & Bois, 1976). In that study, fifty-seven subjects (twenty-nine males and twenty-eight females) segmented each sequence twice, both times in the same instructional condition, in a test-retest reliability study. The test-retest interval was five weeks. Three instructional conditions were employed: Fine-Unit instructions (FU condition), Natural-Unit instructions (NU condition), and Gross- or Large-Unit instructions (LU condition). Sequences were presented in a fixed order, separated by a five-second blank. The content of the sequences was as follows:

Sequence I depicted a man pacing impatiently and intermittently answering a telephone. It was 166 seconds in length.

Sequence II showed a man systematically removing stacks of magazines from a table and shelving them. Length was 157 seconds.

Sequence III showed a woman setting a table with plates and food. Length was 156 seconds.

Sequence IV depicted a man clearing a table cluttered with plates and cups by knocking them off onto the floor. Length was 94 seconds.

Sequence V showed a man systematically building a tower from tinker toys. Length was 154 seconds.

Sequence VI showed a man working at a desk, and occasionally getting up to look in a book on a nearby table. Length was 152 seconds.

Sequence VII depicted a woman making a series of identical stick figures and placing them in a pattern on the floor. Length was 198 seconds.

In constructing these sequences, we attempted to generate sequences that were reasonably diverse, in that they included actions both novel (e.g. VII) and familiar (e.g. III), serious (e.g. VI) and whimsical (e.g. IV), planned (e.g.II) and unplanned (e.g. I). These sequences, however, may not constitute a "representative" sample of behavior. All sequences, for example, employed but one actor. Generalizations from these sequences, therefore, may require qualification on this basis, although it is difficult to specify what a representative sample would be in the absence of a proven taxonomy of behavior. At any rate, the decision to restrict sequences to one actor was a deliberate one in that, at this stage of the research, we wished to avoid complications arising from switching of attention between actors.

One sequence from the Newtson, Engquist, & Bois (1976) study is excluded from the present report. This sequence consisted of a woman dancing to rock music, and it is somewhat different from the others in that it consists of rhythmic movement rather than meaningful, purposive action. Unitization of this sequence was the least reliable of the eight sequences investigated. The primary reason for its exclusion, however, was that number of breakpoints at Large-Unit levels by the criterion employed for the other sequences (see below) were too few for inclusion in the analysis, as a direct

result of its unreliability. Additional analyses of the Fine-Unit and Natural-Unit segmentations including this sequence were conducted, however, and although the sequence itself did not conform to predictions, the obtained pattern of results were identical, and significant differences remained unchanged.

Coding

The position of the actor was coded at each one-second interval of each sequence with the Eshkol-Wachman movement notation (Eshkol, 1973). This notation represents the body as a system of rods moving about the joints; position is specified by spherical coordinates in the "movement sphere" of each limb specifying the joint angle in two dimensions (Eshkol, 1973, p. 8). Codings are recorded by numbers representing a division of the movement sphere into an equal number of parts. A criterion of 45° was adopted for the present analysis; that is, a change in position of a limb such that it became closer to a different point in the movement sphere was required for position to be coded differently, and there were eight such points in each of two planes (vertical and horizontal). For purposes of coding, the body is divided into 15 independently moving parts: left hand, left forearm, left upper arm, right hand, right forearm, right upper arm, head and neck, torso, pelvis, left thigh, left lower leg, left foot, right thigh, right lower leg, and right foot. Two additional features, weight distribution (forward, equal, back) and frontal orientation (in an eight-point circle) are included. This results in seventeen codings of position at each one-second interval.

One other aspect of the coding system deserves mention. Coding of movement is with respect to its pivot joint. Thus, for example, if the actor has his arm fully extended, and raises it while maintaining the position of the forearmand hand relative to the elbow and wrist (i.e., maintaining full extension of the arm), the coding will register a different position for the coding feature upper arm only. Similarly, head and neck movements are coded with respect to the torso. If the actor rotates his upper body, therefore, codings still will differ for torso only. Codings of head and neck position will differ only if the actor turns his head relative to the torso. The result is that this notation system is relatively efficient and non-redundant in its recording of human movement.

To test the distinctive state hypothesis, a position difference index between pairs of points (see below) was computed by comparing the seventeen codings for each point. A difference in coding for a given feature was scored as one, and an identical coding was scored as zero. This yielded an index of position difference between the

points that could range from zero (position the same) to seventeen (position maximally different). For the test of the distinctive change hypothesis, the same index was computed for pairs of points of the different transition types (see below).

The reliability of this change index was assessed by recoding position at all Natural-Unit breakpoints and at an approximately equivalent number of nonbreakpoints. A total 262 points were recoded. The change index was then computed on all successive breakpoint and nonbreakpoint pairs of these points from both sets of codings, and correlated to obtain an estimate of reliability. This value was .84 ($p < .001$), indicating a satisfactory level of reliability for the change index.

Breakpoints were selected from the segmentations obtained in the three instructional conditions in the Newtson, Engquist, & Bois (1976) reliability study. Points were designated as breakpoints at a given level of analysis if the number of unit judgements in an interval ±.5 seconds around that point was one standard deviation above the mean number of judgements per one-second interval in that condition at test or retest. All other points, at one-second intervals, were taken as nonbreakpoints. This resulted in 330 Fine-Unit breakpoints, 133 Natural-Unit breakpoints, and 71 Gross-Unit breakpoints.

Design and Analysis

For the test of the distinctive state hypothesis, Natural-Unit breakpoints for each sequence were paired with nonbreakpoints drawn randomly from the same sequence. These random breakpoint-nonbreakpoint pairs were compared, and the position difference index was computed. Number of data points from each sequence, then, equaled the number of breakpoints in each sequence. To provide a baseline expected value for the difference in position between any two stimulus points drawn at random, for comparison with the previous data, a number of points equal to the number of breakpoints in each sequence were randomly drawn, and then randomly paired with other points from the same sequence. The position difference index was again computed between these random pairs. The distinctive change hypothesis predicts that position difference, on average, between random breakpoint-nonbreakpoint pairs will be greater than position difference between randomly paired stimulus points.

Two additional randomly paired groups were also composed by a similar procedure: 1) random breakpoint pairs, consisting of random pairings of breakpoints within each sequence; and 2) random nonbreakpoint pairs, consisting of random pairings of nonbreakpoints within each sequence. For this latter group, the same number of

random nonbreakpoint pairs was compared as if they were random breakpoint pairs. These two groups provide an indication of the heterogeneity of positions in the two classes of stimulus points.

Data from these four groups were entered into a one-way harmonic means analysis of variance.

For the first test of the distinctive change hypothesis, successive breakpoint-to-breakpoint pairs (B-B transitions) were identified and matched nonbreakpoint pairs were selected on the following basis: the number of one-second intervals separating each B-B transition was determined, and the nonbreakpoint from the center of the next B-B transition was selected (in the case of an even number of intervals, one of the two center points was randomly selected). The next nonbreakpoint at an equivalent number of intervals as spanned by the preceding B-B transition was selected for the matched pairs; if this point was a breakpoint, the starting nonbreakpoint was moved forward or backward in time in the direction that would yield a nonbreakpoint with the least change in the starting point. For this analysis, then, nonbreakpoint pairs were successive, but nonadjacent unless, as occasionally occurred, B-B transitions consisted of adjacent intervals. The change index was then computed for each pair.

For the second analysis, four point-to-point transition types were identified at each level of analysis, and the change computed. The four transition types were B-B transitions (as in the previous analysis, successive in time but not usually adjacent in time); non-breakpoint to succeeding, adjacent breakpoint (N-B transitions); breakpoint to succeeding, adjacent nonbreakpoint (B-N transitions); and nonbreakpoint to succeeding, adjacent nonbreakpoint (N-N transitions).

Finally, the change index was computed for all succeeding stimulus points to provide a baseline change rate over the seven sequences.

A one-way analysis of variance was conducted comparing B-B transitions with the matched N-N transitions on the change index. It was predicted, if the distinctive change hypothesis is correct, that degree of change would be greater for B-B transitions.

Three additional one-way unweighted means analyses of variance (Winer, 1971) were conducted on the four transition types, one for each of the three levels of perceptual analysis. In each analysis there were four levels of the one factor, Transition Type (B-B, B-N, N-B, and N-N). It was predicted that there would be greater change in the stimulus at B-B and N-B transitions than at B-N and N-N transitions at all three levels of analysis.

Results

Comparison of the random B-N pairs with randomly drawn pairs failed to provide support for the distinctive state hypothesis. The random B-N mean was 9.45, which did not differ significantly from the mean for random pairs, 8.97 ($t(406) = .91$, $p < .20$, one-tailed). The mean position difference for random B-B pairs was 10.30, and for random N-N pairs, 9.68. Comparisons of the four means indicated only one significant difference, that between random B-B pairs and random pairs ($t(406) = 2.55$, $p < .05$, two-tailed). This difference might be interpreted as indicating that breakpoints, as a group, are less homogenous than randomly selected points, a result that might be predicted by the distinctive change hypothesis. That is, selection of breakpoints on the basis of their successive change, or difference in position, might be expected to yield, on the whole, a less homogenous set than randomly paired positions. In any event, it is clear that these results do not support the distinctive state hypothesis.

Comparison of the B-B transitions with matched N-N transitions confirmed the first prediction of the distinctive change hypothesis, in the degree of change in the stimulus was significantly greater at B-B transitions ($F(1,262) = 14.27$, $p < .001$). Mean change for B-B transitions was 7.91, and for the matched N-N transitions, 5.17.

Analyses of variance of the four transition types yielded significant effects of Transition Type at Fine-Unit ($F(3,1078) = 34.30$, $p < .001$), Natural-Unit ($F(3,1036) = 41.56$, $p < .001$), and Large-Unit ($F(3,996) = 18.29$, $p < .001$) levels of perceptual analysis. Means are reported in Table Seven, along with results of

---

Insert Table Seven about here

---

t-tests on those means. It was predicted from the distinctive change hypothesis that B-B and N-B means would be greater than B-N and N-N means. As inspection of Table Seven indicates, these predictions were clearly supported at the Fine-Unit and Natural-Unit level of analysis, and received partial support at the Large-Unit level of perceptual analysis. At this level, the predicted difference between N-B and B-N transitions was not obtained. While the means were in the predicted direction, this difference did not approach significance ($t(966) = .46$).

B-N means, in general, were intermediate between N-B means and N-N means, a result that was not predicted. This finding

Table Seven

Mean Number of Feature Changes for Four Transition Types at Three Levels of Perceptual Analysis

Level of Perceptual Analysis	Breakpoint to Breakpoint	Nonbreakpoint to Breakpoint	Breakpoint to Nonbreakpoint	Nonbreakpoint to Nonbreakpoint
Fine Unit	5.51[a]	4.91[b]	4.11[c]	3.14[d]
n	329	150	133	470
Natural Unit	7.91[a]	7.24[a]	5.68[b]	3.41[c]
n	132	89	84	735
Large Unit	7.86[a]	7.15[b]	6.87[b]	3.73[c]
n	70	48	46	836

Note: Means with different superscripts differ within Level of Perceptual Analysis, $p < .05$.

could be due to the fact that, while some changes in the stimulus take more than one second to complete, the pattern of change is unambiguously defined before the movement ends. Movement could thus continue to be large past the point at which the breakpoint itself occurs, but would not be marked as a breakpoint because it has already been registered by the perceiver. For this to account for the large B-N transition mean at large-unit levels, however, would require the additional assumption that this is more the case at this level than at others. While this may indeed be plausible, no evidence for this assertion exists in the present analysis.

Baseline change rate for succeeding, adjacent stimulus points in the seven sequences was 4.30.

Discussion

Results clearly supported the distinctive change hypothesis, while failing to provide evidence for a distinctive state notion. The two hypotheses, it should be noted, are not necessarily incompatible; discrimination of breakpoints could be based upon a combination of the two, or it could be for certain sequences or contexts, distinctive states are relatively more important than distinctive changes. Such was clearly not the case, however, for the range of sequences investigated in the present study.

It might also be argued that the present analysis was insensitive to the distinctive state hypothesis, in that such states are defined not by particular absolute positions of body features, but rather by distinctive configurations of positions that would not be picked up by individual feature comparisons. This same criticism, however, could also be made of the test of the distinctive change hypothesis; distinctive changes could be distinctive configurations of change, and a simple degree of change analysis could thus be expected to be equally insensitive to this hypothesis. Despite these limitations, however, the distinctive change hypothesis was strongly confirmed.

In view of this confirmation, however, one might argue that the assumption that perceptual segmentation is wholly stimulus determined, and hence relatively constant across observers, is indeed justified. The issue here concerns the degree to which the perceiver enters into the definition of action units. One view would be that the perceiver is a relatively passive participant in the process, monitoring the stimulus for a certain degree of change, and interpreting it as it occurs. In this view, which might be termed a "change detection" model, any change above a certain amount would be sufficient for unit formation. Findings that

breakpoints have different informational properties than nonbreakpoints would be seen as due to the inherent organization of the behavior stream, rather than due to active selection of these points according to these properties.

Some evidence against this view is provided by a study reported by Neisser and Becklen (1975). They presented subjects with two optically superimposed sequences of events, in an analogue of a dichotic listening task. Subjects were required to press a key when a significant event occurred in one of the episodes. Neisser and Becklen (1975) found that subjects could readily attend to one episode and ignore the other, even when the two were superimposed. In addition, their subjects rarely noticed the occurrence of "odd events" deliberately inserted into the unattended episodes. Neisser and Becklen (1975) reject explanations of their results in terms of selective filter mechanisms or differing eye-movement patterns, concluding that their results are, "...a direct consequence of skilled perceiving (p. 480)." These results are strong evidence for the active construction of behavior units on the part of the perceiver.

The present approach assumes that the perceiver monitors the ongoing stimulus for particular patterns of change in particular stimulus features, seeing an action as having occurred at those points where such changes occur. It should be emphasized that the seventeen coding features employed in the present analysis are not necessarily those employed by the perceiver. As the coding system is relatively exhaustive with respect to movement, however, at least some of its elements should correspond with the perceptual features actually employed by perceivers. For example, while the system requires separate codings (treats as different elements) of the left hand, left forearm, and left upper arm, these different elements may be employed as a single entity in the perceptual organization of ongoing action.

One further aspect of the present analysis deserves mention. The change index employed in this analysis is a measure of the number of coding features that changed, and is thus only indirectly related to "absolute quantity of stimulus change," however, that might be best defined in a three-dimensional movement sphere.

Despite these limitations, we were curious as to how the monitored perceptual features would be composed from the coding features, and, given the variety of sequences, whether these compositions would be highly variable for sequences of different content. Accordingly, the individual patterns of feature changes in the Natural-Unit B-B transitions were factor analyzed separately for each sequence. Input data consisted of the seventeen zeros and

ones obtained from the comparisons of codings (these were summed to provide the change index). Seventeen variables were input to this analysis, each consisting of a separate coding feature. This data is primarily of descriptive value, given the rather low number of B-B transitions in each sequence relative to the number of input variables.

Results of this analysis, a principal components analysis with varimax rotation, are reported in Table Eight. As inspection of

---

Insert Table Eight about here

---

that table indicates, a clear-cut factor structure was found for each sequence, ranging from four to six factors. Average correlations within and between factors indicate that these factors have a real basis in the data (cf. Nunnally, 1967). In addition, the factor structure for all sequences was a simple one, in that features invariably loaded on one and only one factor. Inspection of the factors themselves suggests face validity. In Sequence I, for example, which portrayed a man pacing back and forth waiting for a series of phone calls, the first factor consists of changes in the features right hand and right forearm, movements of which define answering the phone. Similarly, the second factor in that sequence (see Table Eight) seems to define the actions in response to the phone's ringing, as the actor interrupted his pacing by looking around, turning, and rushing to answer it. In Sequence III, in which a woman set a table with plates and food, the first factor (see Table Eight) seems to correspond well to the actress's stepping up to a table and leaning over to place objects upon it. It is also apparent from these data that the perceptual elements employed consist of quite different coding features in different sequences, suggesting that the perceiver may have considerable flexibility in composing monitored stimulus features.

If these data are meaningful, this would imply that perceivers were actually monitoring between four and six features for change, depending upon the sequence. It would follow, therefore, that a direct test of the feature change model of behavior perception could be made on the basis of a change index composed by treating each factor in the above analysis as a single feature. If at least one of the coding features changed, this could be counted as a change of one; if none of the component coding features changed, this could be counted as zero. The range of this index could be from zero to the number of factors in the given sequence.

Table Eight

Factor Analysis of the Feature Change Structure of B-B Transitions for Seven Behavior Sequences

Sequence	No. of Factors	Eigenvalues	Per Cent Variance	Features Loading >.60	Average r Within Factors	Average Betwee Factor
I	5	5.695	33.5	R. Hand, R. Forearm	.716	.070
		3.409	20.1	Head & Neck, R. Thigh, L. Thigh, L. Forearm, Front		
		1.625	9.6	Pelvis, Wt.		
		1.471	8.7	R. Lower Leg, R. Ft., L. Lower Leg		
		1.168	6.9	L. Upper Arm		
			78.8			
II	5	4.231	28.2	R. Lower Leg, R. Ft., L. Thigh	.478	.052
		2.724	18.2	R. Thigh, L. Lower Leg, L. Ft.		
		2.108	14.1	R. Hand, Head & Neck, Front		
		1.629	10.9	R. Forearm, Torso		
		1.351	9.0	L. Upper Arm, R. Upper Arm		
			80.4			

Table Eight (continued)

Factor Analysis of the Feature Change Structure of

B-B Transitions for Seven Behavior Sequences

Sequence	No. of Factors	Eigenvalues	Per Cent Variance	Features Loading >.60	Average r Within Factors	Average Between Factors
III	6	4.092	25.6	Torso, L. Thigh, L. Lower Leg, L. Ft., Wt.	.430	.149
		2.444	15.3	L. Forearm, L. Upper Arm R. Hand, R. Forearm, R. Upper Arm		
		2.152	13.4	Head & Neck, R. Lower Leg		
		1.643	10.3	L. Hand		
		1.312	8.2	Front		
		1.064	6.7	R. Ft.		
			79.4			
IV	5	4.202	30.0	L. Upper Leg, L. Lower Leg, L. Foot	.759	.133
		2.447	17.5	R. Hand, R. Forearm, Front		

Table Eight (continued)

Factor Analysis of the Feature Change Structure of

B-B Transitions for Seven Behavior Sequences

Sequence	No. of Factors	Eigenvalues	Per Cent Variance	Features Loading >.60	Average r Within Factors	Average Between Factors
IV (cont.)	5	2.000	14.3	L. Hand, L. Forearm	.759	.133
		1.496	10.7	Head & Neck, Torso		
		1.171	8.4	R. Upper Arm, R. Ft.		
V	4	2.958	37.0	L. Hand, L. Forearm, R. Hand, Torso	.623	.082
		1.781	22.3	L. Upper Arm R. Upper Arm		
		1.120	14.0	Head & Neck		
		1.038	13.0	R. Forearm		
			86.2			
VI	6	4.983	29.3	Pelvis, R. Thigh, R. Ft., L. Ft.	.422	.139
		2.925	17.2	L. Upper Arm, Wt.		

Table Eight (continued)

Factor Analysis of the Feature Change Structure of

B-B Transitions for Seven Behavior Sequences

Sequence	No. of Factors	Eigenvalues	Per Cent Variance	Features Loading >.60	Average r Within Factors	Average Between Factors
VI (cont.)	6	1.749	10.3	L. Hand, L. Forearm	.422	.139
		1.673	9.8	R. Upper Arm, R. Lower Leg		
		1.339	7.9	R. Hand, R. Forearm		
		1.199	7.1	Head & Neck, Front		
VII	6	4.982	29.3	Pelvis, R. Upper Leg, R. Ft., L. Ft.	.425	.110
		2.925	17.2	L. Upper Arm, Wt.		
		1.749	10.3	L. Hand, L. Forearm		
		1.673	9.8	R. Upper Arm, R. Lower Leg		
		1.339	7.9	R. Hand, R. Forearm		
		1.199	7.1	Head & Neck, Front		

This analysis was completed according to the above procedure, and the resulting "factor change index" analyzed by a one-way unweighted means analysis of variance comparing the four transition types as before. Results indicated a significant difference between transition types ($F(3,1040) = 30.14$, $p < .001$). Mean number of factor changes at B-B, B-N, N-B, and N-N transitions were 3.69, 2.95, 3.53, and 1.92, respectively. Consistent with the notion that segmentation of the behavior occurred when one of these features changed, the B-B mean was significantly larger than the B-N ($t(1040) = 3.59$, $p < .01$), and the N-B mean was significantly greater than the B-N ($t(1040) = 2.83$, $p < .01$), and N-N means ($t(1040) = 7.88$, $p < .01$). As in the analysis of the raw change index reported earlier, the B-B mean was not significantly different from the N-B mean ($t(1040) = .76$), and the B-N mean was significantly larger than the N-N mean ($t(1040) = 5.04$, $p < .01$).

One further aspect of these data should be noted. The sequences used employed, in general, behavior with a constant theme. If perceivers may vary composition of monitored features from sequence to sequence, they should be able to systematically edit monitored features during ongoing observation as well, as the occurrence of certain events causes the perceiver to anticipate certain other classes of events, or as the theme of the behavior shifts. The composition of non-redundant features for monitoring and the ability to adjust that composition during ongoing observation may be prime components of "skilled perceiving" (cf. Neisser and Becklen, 1975).

The present analyses, then, support the following conclusions: (a) The unit of perception of ongoing behavior comprises the initial perceptual input to processes of person perception; (b) change in the stimulus is a necessary but not sufficient condition for the formation of behavior units; and (c) behavior perception is a feature monitoring process, whereby the perceiver monitors the ongoing stimulus for patterns of change in particular stimulus features, seeing an action as having occurred at those points where changes occur.

Before proceeding on the basis of these data to a formal explication of a model of behavior perception, however, a more direct test of the assumption that actions are defined on a breakpoint to breakpoint basis is necessary. Previous studies have investigated the information properties of either single breakpoints (Experiments Four and Six) or of breakpoint triads (Experiment Five). While the immediately prededing data show there are objective bases for unit formation between successive breakpoints, the action defining properties of breakpoint pairs has not been directly demonstrated. In addition, given our concern with observer skill

and the active, selective processes of the observer, direct evidence that unit formation depends not only upon the change from the previous breakpoint, but also upon the relation of that change and the action it defines to previously perceived actions is essential to formulation of an adequate theory of behavior perception.

Experiment Eight

Accordingly, the procedure employed in Experiment Five was replicated, only with successive breakpoint and nonbreakpoint pairs, rather than the triads employed in that experiment. Subjects viewed seven pairs of breakpoints and nonbreakpoints, judging each pair as to correctness of presentation order and rating their confidence in that judgement. In addition, they rated the degree to which the action portrayed was intelligible, or understandable, and the degree to which it represented a "caused action" as opposed to a random movement.

These data were then analyzed for mean differences, and, where marked heterogeneity of variance was observed, analyzed for systematic differences in variability with an "analysis of variance of spread" (O'Brien, 1976). It was predicted that breakpoint pairs would yield more and/or less variable judgements of ordering, intelligibility, and causal content of action.

While these predictions for pairs follow quite obviously from the strong differences obtained on these measures for breakpoint and nonbreakpoint triads in the Experiment Five study, their confirmation provides both a stronger replication of those findings and baseline data for analyses subsequently reported.

Method

Subjects

Subjects were seventy-nine students (twenty-nine males, fifty females) enrolled in introductory psychology classes at the University of Virginia. Two additional subjects were dropped from the analysis due to failure to complete the measures.

Stimuli

Items were 42 slides (21 breakpoints and 21 nonbreakpoints) drawn from the Newtson and Engquist (1976) study of the informational properties of breakpoints. These slides came from seven 30-second action sequences recorded on 16 mm. black and white film. These consisted of 1) a man nervously leafing through a magazine; 2) a man working on a radio, and smashing it in frustration; 3) a woman cutting out a dress pattern; 4) a man repairing a motorcycle; 5) a woman accidentally spilling a cup of coffee; 6) a man searching

for a lost item in a desk; and 7) a man setting out tools. Items were selected on the basis of the unit markings of twenty subjects instructed to press a button whenever a meaningful action occurred. Slides were made of points where number of marks in an interval ±.5 seconds around that point was one standard deviation above (breakpoints) or one standard deviation below (nonbreakpoints) mean number of marks per one-second interval. An additional constraint on item selection was that, from each sequence, breakpoints and nonbreakpoints alternate. Six slides (three breakpoints and three nonbreakpoints) were selected from each sequence. Two pairs were thus composed from the three points extracted from each sequence.

Apparatus

Slides were presented with a Kodak Carousel slide projector programmed to advance every five seconds, at a distance of approximately eight feet. Projected size was 24 inches on the diagonal.

Measures

For each pair, subjects were asked to: a) judge whether the pair was in correct or incorrect order; b) give a confidence rating for that judgement on a 3-point scale; c) rate the intelligibility of the pair on a nine-point scale from (1) not at all intelligible to (9) very intelligible; d) rate the pair on a nine-point scale on the degree to which the pair represented a "random movement" (1) versus a "caused action" (9); and e) to give a one-sentence description of the two slides.

Procedure

Subjects were run in groups of nine to twelve persons. They were informed that they would see a series of slide pairs, and instructed as to how to complete the measure for each slide set. Each slide in the pair was presented for five seconds; after presentation of the pair, subjects completed the measures, and, when all had finished, the next set was shown. Upon completion of the sets, the purpose of the study was explained and subjects dismissed.

Design and Analysis

Accuracy of order judgements and confidence ratings were combined into a single index by relating the two-valued order judgement (accurate or inaccurate) to the three-point confidence rating. This index ranged from one, assigned to errors made with high confidence, to six, assigned to an accurate judgement with high confidence.

Design for the analysis of variance of this combined confidence and accuracy index and for the measures of intelligibility and caused action was a 2 X 2 repeated measures in which factors were Slide Type (Breakpoints vs. Nonbreakpoints) and Order of Presentation (Correct vs. Incorrect). As there were two pairs of each Slide Type from each sequence containing a common member, four independent groups were run so that no subject saw two pairs of the same Slide Type from each sequence. Data were combined across these groups for the analysis.

Data from the intelligibility and caused action measures were also entered into an analysis of variance of spread (O'Brien, 1976). This technique allows for analysis of differences among the cell variances in factorial analysis of variance designs, testing main effect and interaction hypotheses concerning the variance of a given measure. The basic approach is to use estimates of variability as observations in an analysis of variance. O'Brien (1976) employed both theoretical and Monte Carlo results to evaluate the power and robustness of different estimation techniques in this procedure, demonstrating that two different indices had satisfactory properties, and that both were superior to the usual Z-variance and Box-Scheffe procedures. Of the two indices O'Brien (1976) derived, his "q" values were chosen for use in the present analysis, as it is directly interpretable in terms of an additive model of cell variances. The computational formula employed was:

$$q_{ijk} = \frac{n_{ij}\,(y_{ijk} - \bar{y}_{ij})^2 - s^2_{ij}}{n_{ij} - 2}$$

where n_{ij} is the number of subjects in cell ij of a factorial design, $\bar{y}_{ij}$ is the mean of that cell, and s^2_{ij} is the estimate of the cell variance.

Results

Summary tables for the analysis of variance of the three measures are reported in Table Nine.

Insert Table Nine about here

On the accuracy measure, significant main effects were observed for Slide Type and for Order of Presentation. Mean accuracy

Table Nine

Analyses of Variance of Three Measures of Action Descriptions

		Measure					
		Order Accuracy		Intelligibility		Caused Action	
	df	MS	F	MS	F	MS	F
Source							
Subjects (S)	54	.780		.947		1.375	
Point Type (A)	1	3.510	4.50	1.524	1.61	5.329	3.87
SA	54	.780		.947		1.375	
Order of Presentation (B)		11,818	14.83	3.561	2.22	.710	.41
SB	54	.797		1.603		1.714	
AB	1	.533	.82	.012	.00	.146	.08
SAB	54	.652		2.940		1.740	

for Breakpoint pairs was 3.99, as compared to 3.74 for Nonbreakpoint pairs ($p < .05$). Means for the Correct and Incorrect Presentation Orders were 4.10 and 3.64, respectively, ($p < .01$).

As inspection of Table Nine indicates, mean differences on intelligibility ratings were not significant, although the Breakpoint mean (5.70) was higher than the mean rating for Nonbreakpoint pairs (5.53). Considerable heterogeneity of variance was observed in this analysis, however, and an analysis of variance of spread confirmed that there was significantly greater variability for Nonbreakpoint pairs than for Breakpoint pairs , in that a significant main effect was observed for Point Type in this analysis ($F = 4.36$, $df = 154$, $p < .05$) and mean q-value for Breakpoint pairs was 2.07, and 2.74 for Nonbreakpoint pairs. No other effects attained significance in this analysis.

Breakpoint pairs, consistent with the hypothesis, were rated significantly more as caused actions ($\bar{X} = 5.92$) than were Nonbreakpoint pairs ($\bar{X} = 5.60$, $t = 1.99$, $df = 54$, $p < .05$, one-tailed). Although there was some indication of heterogeneity of variance in this analysis, with greater variance for Nonbreakpoint pairs than Breakpoint pairs, analysis of variance of spread failed to yield a significant main effect for this factor ($F = 1.90$, $df = 1/54$), nor for any others.

Discussion

Results of comparison of breakpoint and nonbreakpoint pairs replicated the findings with point triads in Experiment Five on accuracy of order judgements, although, as could be expected, differences were not as strong. Mean differences in intelligibility ratings, while in the same direction as in the earlier triads investigation, did not attain significance. The extreme (and significant) differences in variability observed in the present investigation on this measure between breakpoint and nonbreakpoint pairs was considerably greater than observed in that study, however.

These results are consistent with the notion that breakpoint pairs define temporal patterns that are the units of behavior processing, in that, as pairs, breakpoints define informationally superior patterns than nonbreakpoints. That is, the relation between successive breakpoints is not only one of greater change (cf. Experiment Seven), but it is also one of greater constraint as to order of occurrence. In other words, the additional degree of change that characterizes the objective difference between breakpoint pairs and nonbreakpoint pairs matched for separation

in time contains additional information as to the ordering of stimulus points. This information, in turn, could result from two different kinds of relationships between degree of change and order information: 1) it could be that greater change, on average, is required to portray non-reversible transformations between stimulus points, or 2) that the more changes between stimulus points the more likely is the occurrence of non-reversible changes. In any case, these data demonstrate that the point-to-point changes that characterize successive breakpoints in action sequences have a consistent informational component that is interpretable by human observers.

Experiment Nine

In this study, judgements of ordering and intelligibility of breakpoint and nonbreakpoint triads were compared to the expected values of these judgements based upon judgements of their component pairs. The issue of concern in this study concerns the basis of breakpoint selection from the behavior stream. The previous data demonstrate that both objective and informational constraints in breakpoint selection do exist between successive breakpoints.

An additional basis for breakpoint selection, however, is the perceiver's ongoing interpretation of the event. That is, unit formation could depend not only upon the change from the previous breakpoint, but also upon the relation of that change and the action it defines to previously defined actions. In this view, the perceiver actively selects breakpoints in support of an overall, ongoing perceptual interpretation.

The issue here is the degree to which a two-action event, as portrayed by three stimulus points, is greater than the sum of its parts. If subjects were segmenting these actions on a simple breakpoint-to-breakpoint basis, one would expect that the accuracy of triad order judgements could be predicted from the accuracy of order judgements of their component pairs. To the degree that subjects selected these points with respect to a larger overall interpretation, however, breakpoint triads should be judged more accurately with respect to order than could be expected on the basis of pairwise judgements.

Method

As the hypothesis concerns the comparison of pair judgements with triad judgements obtained in Experiment Five, some of the data reported in Experiment Nine was employed in the present analysis.

In Experiment Five, triads were presented in correct order (1-2-3) or in a fixed, incorrect order (2-1-3). Measures on 1-2, 2-3, and 2-1 pairs were included in Experiment Nine, and that data was used to compute expected values for this analysis. Data on the additional 1-3 pairs were required, however, for the present analysis, and were obtained concurrently with the data on the other pairs. The constraint was maintained that no subject see two pairs of the same slide type from the same sequence. Data was inadvertently lost due to omission of a 1-3 pair from one of the sequences (Sequence 7), so data from only six of the seven sequences was available for this comparison.

Data and Analysis

Intelligibility ratings for pairs were simply averaged to obtain an expected value for triad ratings. These data were entered into a 2 X 2 X 2 mixed-effects analysis of variance with one between-groups factor and two repeated measures. The between-groups factor was Presentation Format (Pairs vs. Triads), and repeated measures were Slide Type (Breakpoint vs. Nonbreakpoint) and Order of Presentation (Correct vs. Incorrect). There were, then, six observations per cell in this design, as each of the six sequences appeared in each condition. While this may seem rather a low number of observations, it should be kept in mind that each input observation is a mean of ratings, not an individual rating, and thus should provide a very stable estimate of the item properties.

As in Experiment Nine, these data were also entered into an analysis of variance spread (O'Brien, 1976) employing the same factorial design as above. Input-values were computed from the averages of pairs for expected values. If, as predicted, triads contain more information than their independent pairs, these ratings should be less variable for triads than for pairs.

Derivation of order judgement comparisons between the triad data and expected values for triads under the assumption of pairwise independence was done two ways. If subjects were judging the triads by independently evaluating their two component pairs, performance could be a function of one of two indices, depending upon the criterion adopted by subjects. Subjects could have been evaluating each pair, and accepting the triad as correctly ordered if both pairs were judged to be correct. If this were the case, overall probability of a "correct" response to a triad would be expected to be equivalent to the product of the probability of the response "correct" to the component pairs. These expected values were computed for each triad, and, for ease of interpretation, subtracted from one in the Incorrect Order of Presentation

condition so that means would reflect accuracy of performance. The expected values and observed values for triads were entered into a 2 X 2 X 2 mixed-effects analysis of variance in which there were one between-subjects factor, Presentation Format (Pairs vs. Triads), and two repeated measures, Order of Presentation (Correct vs. Incorrect), and Point Type (Breakpoint vs. Nonbreakpoint).

The above criterion assumes, in effect, that subjects were being conservative with respect to judgement of triad ordering as correct, rejecting the triad if one pair was judged as correct and the other as incorrect. Alternately, subjects could have been conservative with respect to judgement of ordering as incorrect, rejecting the triad only if both pairs were judged to be incorrect. Under this assumption, expected value of triad judgements would be equal to the sum of the probabilities of the response "correct" to the pairs minus the product of the two probabilities. These expected values were computed from the pairs data for each triad, and, as before, subtracted from one in the Incorrect Order of Presentation condition so that means would reflect accuracy of performance. These expected values were analyzed with the same design as the previous index.

We shall term the first derived index of expected values a "conjoint" index, as it assumes pair one and pair two of a triad must be identified as "correct" in order for the triad to be judged as "correct." The second derived index of expected triad values we shall term a "disjoint" index, as it assumes that either pair one or pair two of the triad must be identified as "correct" in order for the triad to be judged as "correct." The conjoint criterion assumes that, in general, it is easier to reject a triad as incorrectly ordered than it is to accept it as correctly ordered. The disjoint criterion makes the opposite assumption.

Results

Comparison of the expected value of triads on the intelligibility ratings, as indicated by the average of the intelligibility ratings of the component pairs to the ratings obtained for these triads in Experiment Five, yielded one significant effect, that for Presentation Format ($F = 5.99$, $df = 1/10$, $p < .05$). Mean intelligibility ratings for averaged pairs was 5.82; this mean for the triads was 6.31. Neither the main effect for Point Type ($F = .99$, $df = 1/10$) nor the Presentation Format by Point Type interaction ($F = <1$) was significant in this analysis.

An analysis of variance of spread was conducted on these ratings as in Experiment Nine. Only one effect in this analysis

approached significance, that of Presentation Format ($F = 3.74$, $df = 1/10$, $p < .10$). Mean q-value for pairs was 1.16, and was .366 for triads, in the predicted direction.

The analysis of variance comparing the expected values of accuracy of order judgements on triads to the observed values for those triads for both criteria, conjoint and disjoint, are reported in Table Ten. The significant effects in both analyses of

---

Insert Table Ten about here

Order of Presentation and the Presentation Format by Order of Presentation interactions simply reflect the assumptions made in deriving the expected values. As noted previously, the conjoint criterion implies that it is easier to reject a triad as incorrectly ordered than it is to accept it as correctly ordered, while the disjoint criterion implies the reverse. Consistent with these implications, the Presentation Format by Order of Presentation interaction in both analyses were due to significant differences between the two orders of presentation for expected values only. For the conjoint criterion, mean accuracy was expected to be .456 for Correct Presentation Order, and .763 for Incorrect Presentation Order ($t = 3.39$, $df = 10$, $p < .05$); for the disjoint criterion, mean accuracy was expected to be .740 for Correct Presentation Order and .430 for Incorrect Presentation Order ($t = 3.03$, $df = 10$, $p < .01$). The observed values for triads .625 and .626 for Correct and Incorrect Presentation Orders, respectively, did not differ.

The effects of interest in these analyses are the Point Type by Presentation Format interactions, which were significant in both analyses. Means and the results of t-tests on these means are given in Table Eleven. As inspection of that table indicates, judgemental accuracy of order for breakpoint triads was clearly

---

Insert Table Eleven about here

---

superior to the expected values computed from their component pairs. In addition, observed means for nonbreakpoint triads were less than expected values; for the conjoint criterion this decrement in accuracy approached significance ($t = 1.91$, $df = 10$, $p < .10$, two-tailed), as it did for the disjoint criterion ($t = 1.86$, $df = 10$, $p < .10$, two-tailed.)

TABLE TEN

Analyses of Variance of Accuracy of
Triad Order Judgement

		Criterion for Expected Values			
		Conjoint		Disjoint	
Source	df	MS	F	MS	F
Presentation Format(A)	1	.003	.08	.079	3.16
S(A)	10	.039		.025	
Order of Presentation(B)	1	.284	5.85*	1.150	21.97**
AB	1	.281	5.78*	1.156	22.09**
S(A)B	10	.049		.052	
Point Type(C)	1	.443	11.96**	.293	12.89**
AC	1	.335	9.05*	.494	21.73**
S(A)C	10	.037		.023	
BC	1	.025	.45	.048	1.45
ABC	1	.010	.18	.002	.05
S(A)BC	10	.055		.033	

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

TABLE ELEVEN

Observed and Expected Accuracy
of Order Judgements for Breakpoint
and Nonbreakpoint Triads

		Accuracy Criterion	
Point Type	Observed	Conjoint	Disjoint
Breakpoint	.805[a1]	.622[b]	.521[2]
Nonbreakpoint	.446[b2]	.597[b]	.568[2]

A Theory of Behavior Perception

On the basis of these data, then, one can propose a highly specific model of ongoing behavior perception. An action can be portrayed by a minimum of two successive stimulus points in which at least one common feature has undergone transformation. To the extent that the interpretation of the action depends upon the stimulus content itself, the perceiver has at least two primary sources of information: which features changed, and the nature of the transformation. Some actions may be defined primarily with respect to the latter (e.g. moving an arm), or the reverse (e.g. chasing someone), or some mixture of the two (e.g. raising one's hand). Taken together, these two kinds of information provide a highly useful and flexible basis for behavior perception, and could account for our ability to easily recognize the same actions performed by different persons in different contexts, and our ability to "see" organized action in the movement of figures as abstract as geometric forms (cf. Heider and Simmel, 1944). Some actions, it should also be recognized, may require more than two breakpoints for their definition, and hence capitalize on more complex stimulus properties (e.g. a certain order of changes for specified features, or a particular rhythmic pattern of changes) for their unambiguous definition.

Behavior perception, then, may be viewed as a feature monitoring process. The perceiver monitors some critical set of features (a subset of the available features), segmenting the behavior into parts as one or more of the monitored features change state. Breakpoints, then, are points in the ongoing sequence where a change in state of one or more of the observer's criterial features has occurred.

It is important to note that actions are defined by change in the stimulus, not from the stringing together of a series of discrete states of the stimulus. This definition is critical to the understanding of ongoing action sequences, and has important implications for the way one approaches the perception of meaning in behavior. A meaningful action can only be portrayed by a minimum of two breakpoints in which a common feature has undergone transformation.

This notion is similar to that proposed by the great Russian director and film theorist, Sergei Eisenstein. The primitive way to think about action sequences, Eisenstein (1949) argued, is to see them as composed of a series of building blocks extended linearly in time, with meaning a function of what follows what. The alternative is to view meaning as a function of the successive overlay of images, with meaning defined by the change, or difference

between successive images. In his view, events should be viewed as having a depth dimension in time, consisting of the overlay of successive states, with meaning resulting from the differences between them.

The findings cited above, that the information structure of behavior is highly discontinuous in time, consisting of certain critical information points, is consistent with this view. In addition, it suggests that it may be more profitable to focus upon changes in ongoing behavior per se as the unit of analysis, as opposed to contingencies between the occurrence of specific states.

A number of important implications follow from this model of observation. First, it is clear that behavior perception must impose a short-term memory load on the observer. If an action is defined by a feature change then the state of the feature at time one must be maintained for comparison at time two, when the feature has changed. Given the limited capacity of short-term memory, a critical part of observer skill may be that of feature selection. That is, while many aspects of the ongoing event are changing, some changes may be highly redundant with others, while some may be simply irrelevant. A skilled observer may thus be one who selects the least redundant set of criterial features for perceptual organization of the event, thus insuring maximal information gain from a given observational episode.

In addition, a veteran observer of a given type of event, interaction, or species, may develop a specialized set of predictive features for use in observation, and these may be highly abstract. Observers of mother-child interactions, for example, may employ as a monitored feature the distance between mother and child, and organize their observation (i.e., see something of importance as having occurred) whenever this distance changes markedly. An inexperienced observer could thus view the same event and fail to notice what is readily apparent to the skilled observer, simply because of the difference in criterial features.

Another implication follows from the fact that persons do not monitor stimulus arrays all at once, but employ recursive "scanpaths," moving their eyes around the array in a cyclic manner (cf. Noton and Stark, 1971). Perception of an action depends upon detecting its defining change; the action may thus be perceived as occurring at any time the feature is changed, whether or not the detection occurs at the precise moment the feature underwent the change. If this is true, immediate detection of feature change would be essential for a precise or accurate perception of the order of events.

There seems to be reason to expect that a "span of apprehension" for feature change rates exists, such that features changing too slowly or too rapidly cannot employed for perceptual organization. Consider, for instance, the experience of someone rapidly dribbling a basketball. The cyclic movement of the ball, or perhaps the downward thrusts of the dribbler's hand could, in principle, be employed to organize the sequence bounce by bounce. When the dribbling is rapid, however, we tend to "see," or organize in our mind, the event "dribbling" from the start of it to the end of it. If, however, the dribbler slows down at a continuous rate, there comes a point at which one begins to "see" the event bounce by bounce. The relevant features are now slow enough to be within our span of apprehension for feature changes.

A similar example may be given for slow feature changes. Eibl-Eibesfeldt (1970) cites an instance where a newspaper seller was filmed going about his business. At normal speed, nothing unusual was detected about his behavior. When viewed at fast-motion, however, Eibl-Eibesfeldt reports that the film revealed that the man patrolled a very precisely defined territory, as if he were tethered on a leash (p. 415). Eibl-Eibesfeldt notes that fast-motion film techniques are valuable to ethologists because they "...make visible certain regularities in behavior which normally escape direct observation (p. 415)." Our interpretation would be that the feature changes defining the organization of the behavior--gradual movement around the perimeter--changed too slowly for an observer to employ them to experience the organization that existed in the behavior. It was outside the span of feature change apprehension.

A final set of issues concerns the observer's ability to systematically edit the criterial set of features as observation proceeds. That is, the same features may not be monitored continuously throughout a given observational episode. Given the limits on the number of features monitored, skilled observers may adopt monitoring priorities, such that the appearance of a given feature may cause the observer to cease monitoring another. In addition, shifts in feature monitoring patterns may reflect the observer's dependence upon his causal grammar of the event so that certain feature changes result in systematic shifts to different features, as the occurrence of one action directs the observer to be vigilant for other actions. Insofar as the observer's perceptual grammar corresponds to the actual pattern of feature changes, his efficiency is increased; insofar as it does not, the observer is liable to see causal dependencies in the event that are erroneous.

Phase IV: Direct Investigations of Observer Accuracy

If our reasoning so far is correct, then the perceptual organization of a performance is an important determinant of the information an observer may draw upon in making judgements from his observations.

One basic problem faced by any investigation of observer accuracy is the definition of an adequate criterion against which judgemental accuracy may be scored. In the social judgement literature, such compromise criteria as agreement with estimates of experts, variously defined, or agreement with the average judgement of observers have been employed (Hastorf, Schneider, and Polefka, 1970). Alternately, some studies have asked observers to predict future behavior of the stimulus person, often from unrelated behavior samples, and employed actual performance of the stimulus person as an accuracy criterion.

The literature on objective performance assessment criteria offers no easy solution. As Alluisi (1967) has pointed out, the problem of an adequate criterion for performance on complex meaningful tasks is far from simple, and depends heavily upon decisions and assumptions as to the level of performance to be assessed. While many solutions to the problems have been proposed, none has clearly demonstrated its superiority over a wide range of performance characteristics (cf. Fleishman, 1967).

Because of these difficulties, we decided to employ a class of performance for the present investigation that would minimize uncertainty as to an adequate criterion as much as possible. Subjects viewed a series of films of archers shooting arrows. These stimulus persons were drawn from a physical education course in which long-term records of the objective quality of performance was available. The task met several other criterion, as well, in that it was a skilled motor performance, requiring coordination in the use and manipulation of simple tools. It was brief enough to allow presentation of several performances in a single experimental session, as well as collection of background data. In addition, the performance could be presented from a side view without showing the objective performance outcome. We could thus be confident that observer's judgements were based upon the action itself.

Our object, then, was to provide a stimulus set which definitely contained a performance skill with known, objectively verifiable variations in skill level. This permits verification that performance level can indeed be discriminated from the performance itself

in the absence of more objective performance indices. If one is to employ observer ratings of job performances that do not have clear, immediately evident or conveniently obtainable objective outcome criteria, or for performances where objective outcomes, while evident, are highly variable for even the most highly skilled performers, it is essential that observers be capable of evaluating the action of the rates independent of immediate outcome information.

Given at least a provisional solution to the criterion problem in an accuracy investigation, a second problem concerns the format of the evaluations for computation of accuracy from that criterion. The typical procedure in social judgement research has been to employ rating scales, and such a format was adopted in the present investigation.

As Gage and Cronbach (1955) demonstrated in their classic paper on the analysis of such data, ratings contain two types of accuracy information as well as corresponding sources of bias that may lead to an overestimation of rating accuracy. Because of the complexity of the analysis procedures to cope with these problems, and a corresponding appreciation of the criterion problems inherent in judgements on social dimensions, few systematic investigations have been conducted since Gage and Cronbach's (1955) conceptual analysis. The most complete investigation of accuracy in social perception employing these analytic techniques was reported by Bronfenbrenner, Harding, and Galwey (1958), and their procedures were adopted for the present investigation.

Employing a behavior prediction accuracy criterion, Bronfenbrenner, Harding, and Galwey (1958) provided empirical evidence for the existence of the two accuracy components logically derived by Gage and Cronbach (1955), stereotype accuracy and differential accuracy. Stereotype accuracy refers to the ability to accurately assess the absolute standing on a rating dimension of the group of stimulus persons judged. Differential accuracy refers to the ability to correctly rank a set of stimulus persons on a dimension, regardless of accuracy as to their absolute standing on that dimension. Thus, in a performance evaluation setting, a judge high on stereotype accuracy could correctly discriminate the overall quality of the group, while a judge high on differential accuracy would be able to correctly rank the individual within the group.

In studies of accuracy in person perception, across a wide range of different criterion types, these have been found to be independent, uncorrelated skills. One question of interest in the present investigation is whether this is true of performance

assessment as well. Bronfenbrenner et al. (1958) also reported that the largest component of absolute accuracy in person perception (i.e., raw accuracy scores uncorrected for bias) was due to stereotypic accuracy, with a lesser contribution to overall accuracy of differential accuracy; this question is also of interest in the present context.

In this phase of the research program, three interrelated studies were carried out. The first step consisted of the presentation of observers with skilled and unskilled task performances, obtaining both segmentations and judgements of those performances. Concurrently, a second study of the nature of individual differences in unitization was conducted, so that preliminary evidence as to the relation of observer skill and individual differences could be obtained. A third study was then conducted, employing breakpoints selected on the basis of observer accuracy in the first study, to provide evidence as to the relative importance of information selection as a component of observer skill.

Experiment Ten

The first experiment in this phase was directed at the question of the relation between perceptual organization and observer accuracy. The logic of this investigation was to obtain observer judgements, select the most accurate and least accurate observers, and then to compare their segmentations of the performances. It was recognized that the use of a within groups design has some disadvantages, but given our present state of ignorance as to the nature and basis of observer skill it was impossible to pre-select skilled observers on any objective basis for a criterion-groups type of investigation. This study, therefore, was an exploratory one.

Method

Subjects

Fifty-eight persons (thirty-two males, twenty-six females) were recruited as subjects in the experiment.

Stimuli

Stimuli for the skill assessment consisted of a series of videotapes of ten archers shooting five arrows each. A side view of the archers was presented that did not include the target. Two of these ten served as exemplars for practice trials in the experiment. Of the remaining eight sequences, data from only four,

two at each skill level extreme, are reported here, as preliminary analyses indicated that our observer sample was unable to reliably discriminate the middle skill levels. Sequences ranged in length from 43 to 79 seconds; there was no systematic relation between skill level and duration of the sequence. Sequences were presented in a fixed, randomized order.

Both the scale and the classification of archer skill levels were determined in consultation with the instructor of the archery class from which stimulus persons were drawn. Performances were classified in terms of the proportion of times the given archer could hit a standard target at sixty yards: excellent = 75-100 percent; poor = 0-25 percent. Two intervening skill levels were also identified on this basis, but, as noted, could not be reliably discriminated by our sample of observers. A four point rating scale was used in the experiment.

Apparatus

Videotapes were presented on a 23-inch video monitor. Unit judgements were recorded as in previous experiments.

Procedure

Subjects were run in pairs. To preclude influence between subjects in marking, they were seated at a table divided by a partition, wore headphones, and response boxes were cushioned. Two sample archery sequences were shown, in order to give subjects practice at the task. Subjects segmented each sequence under instructions to "press the button, whenever, in your judgement, a meaningful action occurs; that is, whenever the archer completes a step in shooting an arrow, press the button." After viewing and segmenting each of the two practice sequences, subjects were informed of the skill level of each (2 and 3).

The eight sequences were then presented, and subjects rated each archer on the four-point scale and gave a confidence rating for each judgement. Each archer was rated immediately after his or her performance was viewed.

Measures

Measures consisted of the segmentations of the four archery sequences at skill extremes, ratings of those archers on a four-point scale labeled "very skilled" (1) to "very unskilled" (4), and a corresponding nine-point confidence scale.

Three accuracy indices were derived from these ratings: 1) an absolute accuracy index, incorporating the confidence judgements, that could range from 1 to 18, with 1 equal to an error made with high confidence, and 18 assigned to an accurate judgement made with high confidence; 2) a stereotype accuracy index, consisting of the absolute value of the difference between the subject's mean rating for the four sequences and the true mean on the criterion for the four sequences (2.5); and 3) a differential accuracy index, consisting of a correlation between each subject's four ratings and the corresponding four criterion values. Upon completion of the ratings, a final questionnaire was administered, asking subjects to report (1) their viewing experience with the task; (2) their own experience in performing the task; and (3) self-rated skill level at the task.

Design and Analysis

A full correlation matrix was computed on all measures. In addition, the upper and lower quartiles on the differential and stereotype accuracy indices were identified, and the individual marking patterns for these subjects were summed within each group. Breakpoints for each group were identified, using a criterion of one standard deviation above the average number of marks per interval to identify these points, as in previous studies.

Comparisons of the marking patterns between accuracy and inaccurate groups on each of these two accuracy indices were made by means of a likelihood ratio technique, under assumptions of product binomial sampling. This test statistic was derived specifically for this application,* and its relative power is yet to be determined. It was developed and included because no statistical tests for this type of comparison presently exist, and, if these techniques are to be applied, even roughly appropriate statistical procedures will be helpful in those applications. The test statistic used, T, was as follows:

$$T = 2\{-\sum_{j=1}^{p} (X.j \log_e \hat{\Theta}_j + (N_1 - X.j)\log_e(1-\hat{\Theta}_j)$$
$$+ y.j \log_e \hat{\alpha}_j + (N_2 - y.j)\log_e (1-\hat{\alpha}_j))$$
$$+ \sum_{j=1}^{p} ((X.j + y.j)\log_e \hat{\hat{\tau}}_j + (N_1 + N_2 - X.j - y.j)\log_e(1-\hat{\hat{\tau}}_j))\}$$

* We are grateful to Dr. John Rotondo for providing us with this derivation.

where N_1 is the number of subjects in population one; N_2 is the number of subjects in population two; X.j is the total number of subjects in population one who marked in interval j; y.j is the total number of subjects in population two who marked interval j; $\hat{\theta}_j = \frac{X.j}{N_1}$; $\hat{\alpha}_j = \frac{y.j}{N_2}$; $\hat{\tau}_j = \frac{X.j+y.j}{N_1 + N_2}$; and p is the number of intervals. The null hypothesis is that $\hat{\theta}_j = \hat{\alpha}_j$; the alternative is that there exists some j such that $\hat{\theta}_j \neq \hat{\alpha}_j$. Under the null hypothesis, T is distributed as a chi-square with p degrees of freedom. This value may be converted to a z-score by the formula,

$$z = \frac{(\frac{I}{p})^{1/3} + (\frac{2}{9p}) - 1}{(\frac{2}{9p})^{\frac{1}{2}}}$$

. These unit normal z-values are reported below.

Results and Discussion

Results indicated greater than zero accuracy on all three accuracy indices. Mean absolute accuracy was 8.11, mean stereotype accuracy was .21 (the lower this index the greater the accuracy), and mean differential accuracy was .36. Consistent with previous findings in interpersonal perception, the two accuracy components were independent, as stereotype accuracy correlated only .09 with differential accuracy. This finding may be of considerable practical importance, as it indicates that observer accuracy is a combination of two independent skills, rather than a uni-dimensional skill. While the optimal solution for a given assessment is to have observers with both skills, the importance of the two skill types depends upon the situation. If it is of primary importance that trainees have mastered a task above a certain, absolute level, a stereotypically accurate judge will be adequate, regardless of his standing on differential accuracy. If it is of primary importance to select the best trainees available, regardless of absolute skill level, a differentially accurate judge is most useful.

In this regard, it is of considerable interest that a significant correlation between judge's self-rated skill level and stereotype accuracy was observed ($r = -.38$, $p < .003$). The negative correlation, it should be noted, is due to the fact that a lower score on the stereotype accuracy index indicates greater accuracy. This value was only .18 (n.s.) for differential accuracy. The magnitude of these correlations are probably depressed because of the relatively low number of skilled archers in the sample; mean rated skill level on the same scale as the archers

were rated (1 = very skilled, 4 = very unskilled) was 3.19. They suggest, however, that experience with a given skill more readily confers accuracy in absolute judgement of skill level, rather than differential sensitivity.

Maier (1976) notes that the Army has decided to shift from paper and pencil testing to "criterion-referenced performance testing." Part of the new tests involve observer ratings of task performances; observers rate the trainee in terms of whether his performance is above an absolute level of proficiency. Maier (1976) reports that a preliminary decision has been made to select observers on the basis of self-rated skill level at the task. These data indicate that that decision was a sound one.

Contrary to findings on accuracy in social perception (cf. Bronfenbrenner et al. (1958)), the bulk of overall accuracy in these data was accounted for by differential accuracy. Differential accuracy correlated .42 ($p < .001$) with absolute accuracy, while this value was only -.15 (n.s.) for stereotype accuracy. This finding is reasonable in view of the relation between stereotype accuracy and the observer's task competence, and the low level of task skill in the sample. In relation to Bronfenbrenner et al.'s (1958) findings that the largest accuracy component in normal social perception in interactions was due to stereotypic accuracy, this finding again suggests that observer expertise in the task primarily enhances accuracy in the absolute level of ratings.

Subjects were selected who fell into the upper and lower quartiles of each accuracy measure, and their segmentations of the stimulus sequences were compared by the method previously described. According to this test, segmentations of stereotypically accurate subjects differed significantly from segmentations of stereotypically inaccurate subjects ($Z = 2.43$, $p < .01$); segmentations of differentially accurate subjects also tended to differ from those of differentially inaccurate subjects ($Z = 1.11$, $p < .14$). Further examination of the segmentations of differentially accurate and inaccurate subjects indicated that these two groups differed strongly in their segmentations of the least skilled performances ($Z = 2.23$, $p < .05$), but not in the segmentations of the most skilled performances ($Z = 1.39$, $p < .10$).

These differences, it should be noted, were in patterns of segmentation. Simple number of units marked correlated -.01 with absolute accuracy, .06 with stereotype accuracy, and .03 with differential accuracy. As these subjects subsequently took part in a study of individual differences in unitization (Experiment

Twelve, see below), it was possible to relate those findings to observer accuracy in the present data. Unitization range, defined as the difference between number of units marked under fine-unit instructions minus the number of units marked under the large-unit instructions for the same behavior sequence correlated .28 ($p < .05$) with differential accuracy. This suggests that those persons with the greatest flexibility in perceptual organization are more able to discriminate individual differences in performance skill. Range was unrelated to stereotype accuracy ($r = .08$).

The relation between perceptual organization and observer accuracy in the present data supports the notion that an important part of both types of observer skill is the perceptual selection of good information. It remains to be demonstrated, however, that these segmentation differences played a causal role in judgemental accuracy. That is, while pattern of perceptual organization would appear to discriminate between accurate and inaccurate observers, the differences in segmentations could be the result of incidental, correlated differences between accurate and inaccurate observers, and not of causal significance.

This issue is of practical importance as well, in that it has implications for observer skill training. If information selection, as indicated by perceptual organization, is an important causal factor in observer accuracy, then observer training might best include direct perceptual practice with segmentations of skilled observers.

One straightforward means of investigating this issue might be to select the breakpoints of the high and low stereotype and differential accuracy groups, photograph them, and mount them as slides. These slides could then be presented to four independent groups of judges, who would then be asked to rate the performances on the same scales. Differences between the high and low accuracy groups, if obtained, might then be taken as evidence for the causal role of the segmentations in observer accuracy.

One problem with this approach, however, is that it relies on the assumption that the segmentations of the inaccurate groups, and, presumably, their interpretations, are as homogeneous as those of the accurate groups. A more reasonable assumption would be that, while there is a uniform perceptual organization for accurate judges, in that they all converge to the same judgement, observers may be inaccurate for a wider variety of reasons. For example, stereotypically inaccurate judges were those who either over-estimated or under-estimated the skill level of the group. Similarly, differentially inaccurate judges were those who either

ranked the stimulus persons in a manner uncorrelated with the criterion, or those who produced rankings inversely related to the criterion. Both groups could include those who were simply confused, and failed to come up with a consistent perceptual interpretation.

These problems compromise the interpretation of the segmentation differences obtained between skilled and unskilled observers, as well. These differences could have been an artifact of comparison of a group selected on the basis of a uniform interpretation (the accurate groups) with the segmentations of a group with a variety of interpretations (the inaccurate groups).

With respect to the interpretation of the simple decoding study outlined above, furthermore, differences in decoding accuracy could simply reflect a difference in coherence of the accurate segmentations, based upon a uniform interpretation, and inaccurate segmentations composed of combined segmentations of a variety of interpretations. Failure to find a difference in decoding accuracy, furthermore, could result if the combination of two entirely different inaccurate segmentations cancelled out each other's deficiencies. For example, combining the segmentation of a judge who over-estimated skill levels with the segmentation of a judge who under-estimated skill levels might provide the basis for an accurate estimation.

One further complication is introduced by the fact that the right information may be a necessary, but not sufficient condition for judgemental accuracy. That is, a skilled observer might be unable to make an accurate judgement given unsuitable information, but an unskilled observer might also be unable to make an accurate judgement whether he has the correct information or not.

Experiment Eleven

In light of these problems, the following procedure was adopted in order to provide evidence for the causal role of behavior segmentation in observer accuracy. Breakpoints from the stereotypically accurate group in the previous study were identified and mounted as slides, as were the breakpoints from the differentially accurate group. The resulting two slide sets were then presented, separately, to two independent groups of judges. These judges then rated the skill level of each archer, as in the previous study. Both stereotypic and differential accuracy indices were then computed and compared to the average levels of accuracy on these indices obtained in the previous study from judges viewing the continuous sequences.

The logic of these comparisons is as follows. Both groups of judges, those who viewed the continuous sequences and those who viewed the segmentations only, may be assumed to be, on average, of equal skill levels with respect to their ability to integrate information concerning the task. The highly accurate groups who provided the basis for the selection of breakpoints could have been superior both in information selection and in information integration skills. If the breakpoints of the accurate groups do contain the informational base of the judgements, then we have provided the decoding groups with the information necessary for accurate judgements. If this information is not only necessary, but also sufficient, the decoding groups would be as accurate as the groups from whom the breakpoints were selected. If this information is necessary, but not sufficient, the decoding groups should do as well as the groups who saw the continuous sequences. If, however, the segmentations of the highly accurate groups do not contain the informational base of the judgements, then subjects viewing the continuous sequences should be more accurate than subjects viewing the segmentations only, and our initial hypothesis in this phase of the research that the perceptual segmentation of a performance determines the information an observer may draw on in making his judgement is clearly disconfirmed.

The hypothesis, then, may be tested twice, on both differential and stereotypic accuracy skills. Our prediction is a null one, however, and thus may provide at best only indirect evidence for the validity of the hypothesis. More direct evidence may be provided by taking advantage of the independence of the two types of observer accuracy. To the extent that the two accuracy skills are stimulus information selection skills, and not information integration skills, subjects viewing breakpoints from judges high on only one of the skills should achieve their level of absolute accuracy primarily on the basis of that type of accuracy. That is, level of absolute accuracy should correlate with stereotypic accuracy for subjects viewing the breakpoints of stereotypically accurate judges; similarly, level of absolute accuracy should correlate with differential accuracy for subjects viewing the breakpoints of differentially accurate judges.

These predictions depend upon the assumption that the two independent components of accuracy are stimulus based. For this assumption to be true, the hypothesis that the perceptual segmentation determines the informational base of observer judgements must also be true. Confirmation of these predictions, then, will provide the evidence needed for confirmation of that hypothesis, although disconfirmation will not rule it out.

The result, then, is that we have the possibility of clearly disconfirming the hypothesis that perceptual segmentation constitutes the basis of observer judgements, and the possibility of clearly confirming that hypothesis.

Despite the ambiguities noted above in interpretation of the segmentations of the inaccurate groups, these segmentations were also included in the present study.

Method

Subjects

Subjects were sixty-three undergraduates at the University of Virginia.

Stimuli

Stimuli were slides of breakpoints identified from the markings of the upper and lower quartile groups on the stereotypic and differential accuracy measures. Criterion for breakpoint selection was that any point above the mean number of marks per interval was taken as a breakpoint. This resulted in 84 breakpoints from the stereotypically accurate group. Slides were prepared of the differential accuracy points, and the time intervals between them were recorded. Points from the stereotypically accurate group were recorded as stills on videotape and presented to subjects on a video monitor. In both formats, each breakpoint was presented and maintained for the actual length of time between it and the occurrence of the next breakpoint, and then the next breakpoint was shown.

Breakpoints from the lower quartile differentially accurate and stereotypically accurate groups were obtained in an identical manner, and prepared as slides.

Procedure

Subjects were run in groups ranging from one to seven. Subjects were instructed that their task would be to view a series of archers, to rate the skill level of each, and to give a confidence judgement for each rating. They were then shown the same two practice videotapes that were used in the previous experiment. With the exception of the marking instructions, instructions to this point were exactly the same as in the previous experiment. Subjects were then informed that, rather than continuous videotapes, they would view a series of stills, but that the duration of each still

picture represented the actual time it took for the actor to change from one position to the next.

Measures

The three accuracy indices, absolute accuracy, stereotype accuracy, and differential accuracy were computed exactly as in the previous experiment.

Results and Discussion

Mean differential accuracy of subjects who viewed the continuous videotapes in the previous experiment was .52; subjects who viewed the breakpoints of the differentially accurate group averaged .51 on this index (t (78) = .09, n.s.). The upper quartile group, selected from the first experiment, had a differential accuracy rating of .90, which was significantly more accurate than ratings of the decoding group (t (934) = 2.98, $p < .01$). Mean stereotype accuracy of subjects who viewed the continuous videotapes was .31; subjects viewing the breakpoints of the stereotypically accurate group averaged .455 (less accurate on this index, for which a lower value indicates greater accuracy), but this difference was not statistically significant ($t(67) = 1.59$, $p < .20$). The upper quartile group on this index, from whom these breakpoints were selected, had a stereotype accuracy mean of .27, which was nearly significantly more accurate than the decoding group ($t(24) = 1.81$, $p < .10$).

With respect to the first set of predictions then, these data do not confirm that perceptual segmentation is unrelated to judgemental accuracy, in that, on both indices, the decoding groups were not significantly less accurate than subjects viewing the continuous sequences themselves. The pattern of results are clearly consistent with the proposition that the perceptual segmentation of accurate observers, as represented by breakpoints alone, is a necessary but not sufficient condition for judgemental accuracy.

More direct evidence for this proposition is provided by the correlations between absolute accuracy and the two independent accuracy indices in the decoding groups. In addition to stimulus-based discrimination, it will be recalled, absolute accuracy scores contain additional components of bias (cf. Gage & Cronbach, 1955). The two accuracy indices were specifically derived to control for these biases. Insofar as absolute accuracy of judgement is based upon stimulus information, however, this index should correlate with one of the two uncontaminated indices, and thus represent the degree to which absolute accuracy is accounted for by each skill (Bronfenbrenner et al., 1958).

For subjects viewing breakpoints from the stereotypically accurate group, absolute accuracy correlated -.56 with stereotype accuracy ($p < .07$; note that the inverse relation follows from the nature of this index), and only .08 (n.s.) with differential accuracy. For subjects viewing breakpoints from the differentially accurate group, absolute accuracy correlated .59 ($p < .001$) with differential accuracy, and only .08 (n.s.) with stereotype accuracy. Intercorrelations of the two indices within both decoding groups again indicated independence of the two accuracy types, correlating -.04 and .06 within these conditions.

The results confirm that the two independent components of judgemental accuracy are independently based in the stimulus, in that they decoded independently, and hence, by implication, directly confirm that the perceptual segmentation of a task performance determines the informational base of observer judgements.

Additional evidence for this proposition comes from a comparison of mean differential accuracy of decoding groups who viewed the breakpoints of stereotypic or differentially accurate judges. These two breakpoint selection groups differed strongly on level of differential accuracy, with the differential accuracy selection group significantly more accurate ($t(27) = 4.28$, $p < .01$). This accuracy difference decoded quite strongly, as well, with subjects viewing breakpoints of differentially accurate judges significantly more accurate on this index than subjects viewing breakpoints of stereotypically accurate judges ($t(931) = 2.11$, $p < .05$).

A similar comparison on the stereotype accuracy index was not informative, as the differences on this index between selection groups did not attain significance ($t(27) = .06$), nor did the two decoding groups differ significantly ($t(31) = 1.46$, $p < .20$). This problem reflects a consistent pattern of weak results on the stereotypic accuracy measure throughout these investigations that probably reflect the constraints of the present procedures. In particular, with only a four-point rating scale and the concomitant distribution of skill levels such that the criterion and scale midpoints coincide, this index was probably quite insensitive in these studies. Future investigations would do well to avoid these constraints.

Finally, comparisons of the absolute accuracy of subjects viewing breakpoints of accurate vs. inaccurate judges indicated that subjects viewing only the breakpoints of accurate judges were more accurate in their judgements ($\bar{X} = 8.76$) than judges viewing breakpoints of inaccurate judges ($\bar{X} = 7.70$, $t(59) = 2.48$, $p < .01$). As noted above, interpretation of findings employing segmentations of inaccurate judges is ambiguous.

These data provide strong confirmation that observer information selection is a prime component of observer accuracy, and imply that observer skill training in segmentation may be an important part of any program to enhance observer accuracy.

Experiment Twelve

One final question of both pragmatic and theoretical interest concerns the nature of individual differences in perceptual organization. If, as the previous data suggest, observer skill depends upon highly specific differences in perceptual organization, then skill training in observation should be highly feasible. This would be consistent with Barker and Wright's (1955) conclusion from their attempts to maintain reliability of level of organization across observers. Alternately, observer skill could depend upon more enduring cognitive and affective structures, such as personality variables. If this is the case, it could be more profitable to pre-select observers according to those characteristics that are accuracy-related than to undertake observer skill training.

The relation between this perceptual system and personality variables is of theoretical interest as well. At the outset of this project, our preliminary evidence suggested that behavior perception is a process that occurs at a low cognitive level; consistent with this interpretation, one would anticipate that no strong relationships with global personality measures should be found.

In terms of rounding out our evidence as to the nature of this process, however, some evidence as to the nature of individual variation in perceptual organization is necessary. One question concerns the nature of individual differences in the "normal behavior perspective," to use Barker and Wright's terminology (1955). Do individuals differ systematically in the size, or extent, of their behavior perspective such that some persons are able to analyze behavior at extremely fine or extremely large levels, while others operate within a more restricted range? Or do persons tend to differ in the overall level of their range of analysis, such that some persons are able to analyze very finely, but not at very large levels, or vice versa? Do these characteristics relate in any systematic fashion to other individual differences?

Method

Subjects

The fifty-eight subjects who participated in Experiment Ten also participated in the present study.

Stimuli

The stimulus tape consisted of a 154-second sequence depicting a man constructing a tower from blocks and sticks (Sequence VI, Experiment Three). This sequence was chosen because it provided a clear discrimination between the three instructional levels of analysis (fine-unit, large-unit, and natural-unit) with a high degree of reliability at each level.

Procedure

Assessment of observer segmentation characteristics was conducted following the subjects' participation in Experiment Ten. Subjects were given a five-minute rest period after completion of that experiment. All subjects segmented the sequence first under instructions to segment the behavior into "whatever actions seem natural and meaningful to you"(Natural-Unit instructions). Marking instructions for the second and third viewing were varied, so that half of the subjects were instructed to mark the sequence into "the smallest actions that seem natural and meaningful to you," (Fine-Unit instructions) and then, on the third viewing, to segment the sequence into "the largest units that seem natural and meaningful to you" (Large-Unit instructions). The other half of the subjects received these same instructions in reverse order (Large-Unit, then Fine-Unit).

Upon completion of the final viewing, subjects were requested to complete the short form of a personality inventory, the Personality Research Form. Subjects were informed prior to completing the instrument that a) its completion was not mandatory, and b) their name was not to be put on the test, as their responses were completely anonymous. That is, each subject's data set was given a number so that, while data could be analyzed for systematic relations with the prior accuracy assessment, no link between the subjects' name and their data could be established, thus assuring complete privacy for this personal information. All subjects agreed to complete the inventory.

Measures

Measures consisted of the three segmentations at each of the three instructional levels. Direct measures, then, consisted of the number of units marked under Fine-Unit (FU), Natural-Unit (NU), and Large-Unit (LU) instructions. In addition, two further measures of observer characteristics were derived: Range, consisting of the number of units under FU instructions minus the number of units under LU instructions (i.e. Range = FU - LU), and a "natural unitization tendency" score (NUTS), consisting of the number of units

at NU instructions minus number of units at LU instructions, divided by the range (i.e. $\frac{NU - LU}{FU - CU}$). This "NUTS" index provides a measure of the place in the observer's range where the observer chooses to operate under NU instructions relative to his range. Upon completion of the segmentation tasks, subjects completed the Personality Research Form, Form A (Jackson, 1967). This form was chosen because it has the best psychometric properties available in current personality tests and is especially appropriate to this sample.

Design and Analysis

Full correlation matrices were obtained for all measures. In addition, number of units employed in the three instruction conditions were analyzed in a 2 X 2 X 3 mixed effects analysis of variance in which factors were Sex, Order of Instructions (NU - FU - LU vs. NU - LU - FU), and a repeated measure, Instructional Set (NU vs. FU vs. LU).

Results

Analysis of variance yielded only one significant effect, that for Instructional Set ($F(2,108) = 99.72$, $p < .001$). No other effects or interactions were significant. Mean number of units employed in the three instructional conditions were 30.26, 49.53, and 12.16 for the Natural-Unit, Fine-Unit, and Large-Unit conditions, respectively.

Intercorrelations of the different measures of observer segmentation characteristics are reported in Table Twelve. As

---

Insert Table Twelve about here

---

inspection of that table indicates, there was a positive and highly significant correlation between number of units employed at Fine-Unit levels and number of units employed at Large-Unit levels. This relation indicates that persons tend to differ in the overall level of their range, such that those persons who analyze most finely are least able to analyze in large units. Results on the "NUTS" index indicated that there is no systematic relation between where in the range an individual typically operates and the size of the range. Mean value for this index was .51 (SD = 1.22), indicating that, in general, persons operate precisely in the middle of their available range.

TABLE TWELVE

Experiment Twelve

Intercorrelations of Five Observer Segmentation Characteristics

	Characteristic				
	No. Fine Units	No. Large Units	No. Natural Units	Range	NUTS
No. Fine Units	1.00	.471**	.799**	.919**	-.053
No. Large Units		1.00	.572**	.084	-.056
No. Natural Units			1.00	.647**	.264*
Range				1.00	-.035
NUTS					1.00

* $p < .10$

** $p < .001$

Only four correlations between the fifteen personality dimensions of the Personality Research Form and the five segmentation characteristics even approached significance, about what would be expected by chance. This is consistent with interpretation of the process of perceptual organization of behavior as a highly specific cognitive perceptual skill.

With respect to the two types of observer accuracy, analysis indicated an absolute lack of relation between personality dimensions and observer accuracy of either type. There was, as noted previously, a significant relation between range size and differential accuracy ($r = .28$, $p < .03$), although not between range size and stereotype accuracy. No other significant relations between accuracy and unitization properties of the observer were obtained.

Lack of a relationship between this process and global individual differences, however, does not rule out the possibility that this process may mediate, and hence be biased by, interactions between behavior content and personality or attitudinal predispositions.

For example, we have not found systematic sex differences as a function of either sex of observer or sex of stimulus person. Dr. Kay Deaux, of Purdue University, has, however, investigated these variables in relation to behavior segmentation in somewhat more detail. Deaux (personal communication) varied sex of observer and sex of stimulus person, and, as we have, found no overall difference on either variable. In addition, however, she included Bem's Androgeny Scale - an attitudinal measure of the degree to which persons accept conventional, traditional sex-role definitions. She found that both males and females who are highly sex-typed segmented male task behavior more finely than they segmented female task behavior (the task performance was identical for both sexes of stimulus person).

To the extent that segmentation differences can alter judgements of performance, a constant bias would be produced by sex-typed raters; this bias would not be countered by matching raters to ratees by sex, furthermore, as highly sex-typed women show an identical bias to that of highly sex-typed men.

It may be, therefore, that personality-situation interactions are mediated through effects on perceptual organization of events, as, when the content of an event is relevant to a personality dimension of the perceiver, his interpretations of that event undergo systematic biases.

Summary and Conclusions

The overall goal of this program of research was to develop a theoretical model of the human observer with definite implications for observer skill. In order to progress towards this goal it was first necessary to identify, at least on a preliminary basis, the level at which observational processes occur. Our evidence indicated that our phenomena are products of low level perceptual-attentive processes. The second step was to verify the phenomenal reality of the units discriminated at the perceptual level. Our evidence indicated that such units are subjectively real units of experience of ongoing behavior, and also offered clues as to the nature of their informational base. Of critical importance was the implication that the perceptual structure of behavior understanding rests upon the discrimination of intermittently-occurring points of definition in the behavior stream. This possibility formed the basis of the next step in the investigations, an inquiry into the objective structure of behavior as a stimulus, and its relation to the phenomenal units by which actions are comprehended. At last, a general conceptual framework that could reasonably account for the phenomenon and evidence at hand began to emerge. While far from complete, the model of behavior perception as a feature monitoring process can serve as a basis for the generation and evaluation of specific hypotheses about specific phenomenon in ongoing observation.

Given this conceptual framework, the final task was to verify that it could indeed serve as an operationalization of observer skill. Despite the complexities of accuracy research, supportive evidence for this application of the conceptual framework was obtained. Such evidence as was gained, however, suggested that, while the model succeeded in operationalizing the information selection basis of observer skill - a necessary basis for accurate observation - further processes of information integration must be also accounted for in a complete description of skilled observation.

In this connection, it is well to keep in mind that an ongoing perceptual process will almost necessarily be a cyclic one, with constant feedback between processes of information selection and interpretation. The perceiver constantly works out interpretations and seeks information in order to generate new interpretations. Perception, then, is not a matter of passive interpretation of stimulus information, nor is perceptual skill simply a matter of the precision or efficiency of information integration. Perception, and behavior perception in particular, is a process of active interaction between the perceiver and the potential of his world for information.

With respect to the pragmatic concerns that motivated the support of this project, one final point is in order. The conclusions and implications of this research are highly speculative, and at most can be said to provide a heuristic framework within which particular, specific applications may be approached. It may well be the case that the methods we have developed in the course of this research for addressing questions of observer skill in performance assessment will be of greater value than any of the conclusions as to the nature of observational processes we have put forward.

BIBLIOGRAPHY

Alluisi, E.A. Methodology in the use of synthetic tasks to assess a complex performance. Human Factors, 9, 1974.

Altman, J. Observational study of behavior: Sampling methods. Behavior, Vol. XLIX, 1974.

Barker, R.G. & Wright, H.F. Midwest and its children. Hamden, Connecticut: Archon Books, 1955.

Barker, R.G. (Ed.) The stream of behavior. New York: Appleton-Century-Crofts, 1963.

Bieri, J., Atkins, A.L., Briar, S., Leaman, R.L., Miller, H., & Tripodi, T. Clincal and social judgement: The discrimination of behavioral information. New York: Wiley, 1966.

Bronfenbrenner, U., Harding, J. & Gallwey, M. The measurement of skill in social perception. In D. McClelland, A.L. Baldwin, U. Bronfenbrenner, & F. Shodtbeck (Eds. Talent and society, Princeton, N.J.: D. Van Nostrand Co., 1958.

Dickman, H.R. The perception of behavioral units. In R.G. Barker (Ed.) The stream of behavior, New York: Appleton-Century-Crofts, 1963.

Easterbrook, J.A. The effect of emotion on cue utilization and the organization of behavior. Psychological Review, 66, 1959, 183-201.

Eibl-Eibesfeldt, I. Ethology: The biology of behavior. Translated by E. Klinghammer; New York: Holt, Rinehart, & Winston, 1970.

Eisenstein, S. The film form. New York: Harcourt Brace & Co., 1949.

Fleishman, E.A. Performance assessment based on an empirically derived taxonomy. Human Factors, 9, 1974.

Fodor, J.A. & Bever, T G. The psychological reality of linguistic segments. Journal of Verbal Learning and Verbal Behavior, 4, 1965, 414-420.

Frey, J., & Newtson, D. Differential attribution in an unequal power situation: Biased inference or biased input? Proceedings, 81st Annual Convention of the American Psychological Association, 1973.

Gage, N.L. & Cronbach, L.J. Conceptual and methodological problems in interpersonal perception. Psychological Review, 62, 1965, 411-422.

Glass, D.C. & Singer, J.E. Urban stress: Experiments on noise and social stressors. New York: Academic Press, 1965.

Hastorf, A.H., Schneider, D.J., & Polefka, J. Person perception. Reading, Massachusetts: Addison-Wesley, 1970.

Heider, F. & Simmel, M. An experimental study of apparent behavior. American Journal of Psychology, 57, 1944, 243-259.

Jackson, D.N. Personality Research Form. Goshen, N.Y.: Research Psychologists Press, 1967.

Jenkins, J.J., Wald, J., & Pittenger, J.B. Apprehending pictorial events: An instance of psychological cohesion. In C.W. Savage (Ed.), Minnesota studies in the philosophy of science, Vol. 9, in press.

Jones, E.E. & Davis, K.E. From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), Advances in experimental social psychology, Vol. 2, New York: Academic Press, 1965.

Kahneman, D. Attention and effort. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Kelley, H.H. Attribution theory in social psychology. Nebraska symposium on motivation, 15, 1967, 192-238.

Leventhal, H. Findings and theory in the study of fear communications. In L. Berkowitz (Ed.), Advances in experimental social psychology, Vol. 5, New York: Academic Press, 1970.

Lyon, J. The perception of human actions. Journal of General Psychology, 54, 1965, 45-55.

Maier, M. Performance-based testing in the U.S. Army. American Psychological Association, 1976.

Miller, G.A. & Johnson-Laird, P.N. Language and perception. Cambridge, Massachusetts: Belknap/Harvard University Press, 1976.

Neisser, U. & Becklen, R. Selective looking: Attending to visually specified events. Cognitive Psychology, 7, 1975, 480-494.

Neisser, U. Perceiving, anticipating and imagining. Proceedings, American Psychological Association, 1975.

Newtson, D. Attribution and the unit of perception of ongoing behavior. Journal of Personality and Social Psychology, 28, 1973, 28-38.

Newtson, D. Foundations of attribution: The perception of ongoing behavior. In J. Harvey, W. Ickes, & R. Kidd (Eds.), New directions in attribution research, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

Newtson, D., Engquist, G., & Bois, J. The reliability of a measure of behavior perception. JSAS Catalog of Selected Documents in Psychology, 6, 1976, 5, (MS 1173).

Noton, D. & Stark, L. Scanpaths in saccadic eye movements while viewing and recognizing patterns. Vision Research, 11, 1971, 929-942.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

O'Brien, R.G. Robust techniques for testing homogeneity of variance hypotheses in factorial designs. Joint meeting of Psychometric Society and Mathematical Psychology Group, 1976.

Sarbin, T.R., Taft, R., & Bailey, D.E. Clinical inference and cognitive theory. New York: Holt, Rinehart, & Winston, 1960.

Siegal, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Snodgrass, J.G. Psychophysics. In B. Scharf (Ed.), Experimental sensory psychology, Glenview, Illinois: Scott Foresman & Co., 1975.

Snodgrass, J.G., Volvovitz, R.J., & Walfish, E.R. Recognition memory for words, pictures, and words and pictures. Psychonomic Science, 27, 1972, 345-347.

Snyder, M. Self-monitoring and expressive behavior. Journal of Personality and Social Psychology, 30, 1974, 526-537.

Wiggins, N., Hoffman, & Taber. Types of judges and cue utilization. Journal of Personality and Social Psychology, 11, 1969,

Zadny, J., & Gerard, H.B. Attributed intention and informational selectivity. Journal of Experimental Social Psychology, 10, 1974, 34-52.

Chronological Bibliography of Publications

Resulting from Grant to Date

Newtson, D. Foundations of attribution: The perception of ongoing behavior In J. Harvey, W. Ickes, & W. Kidd (Eds.) New directions in attribution research. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

Newtson, D., Engquist, G., & Bois, J. The reliability of a measure of behavior perception. JSAS, Catalogue of Selected Documents in Psychology, 1976, 6, 5, (MS 1173).

Newtson, D. The process of behavior observation. Journal of Human Movement Studies, 1976, 2, 114-122.

Newtson, D. & Engquist, G. The perceptual organization of ongoing behavior. Journal of Experimental Social Psychology, 1976, 12, 436-450

Graduate Students supported by this grant: (includes those who received hourly pay for specific small tasks.)

Gretchen Engquist (M.A.)
Jeffrey Friedman (M.A.)
Robert Thibadeau (Ph.D.)
Rick Rindner
Robert Campbell
Carol Toris
Christopher Massad (M.A.)
Mark Krischik