DOCUMENT RESUME

ED 160 918 CG 012 850

AUTHOR Farkas, Gary M.; Tharp, Roland G.
TITLE Observation Procedure, Observation Gender, and Behavior Valence as Determinants of Sampling Error in a Behavior Assessment Analogue.
PUB DATE 78
NOTE 29p.; Paper presented at the Annual Meeting of the Western Psychological Association (58th, San Francisco, California, April 19-22, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Behavioral Science Research; College Students; Expectation; Group Behavior; *Observation; Perception; Research Methodology; Research Projects; *Sampling; *Sex Differences
IDENTIFIERS *Sampling Error

ABSTRACT

Several factors thought to influence the representativeness of behavioral assessment data were examined in an analogue study employing a multifactorial design. Systematic and unsystematic methods of observing group behavior were investigated using 18 male and 18 female observers. Additionally, valence properties of the observed behaviors were inspected. Subjects' assessments of a videotape were compared to a criterion code which defined the population of behaviors. Results indicated that systematic observation procedures were more accurate than unsystematic procedures, though this factor interacted with gender of observer and valence of behavior. Additionally, males tended to sample more representatively than females. A third finding indicated that the negatively valenced behavior was overestimated, while the neutral and positively valenced behaviors were accurately assessed. (Author)

ED160918

Observation Procedure, Observer Gender, and Behavior Valence as Determinants of Sampling Error in a Behavior Assessment Analogue

Gary M. Farkas and Roland G. Tharp

University of Hawaii at Manoa

and

Kamehameha Early Education Program

Address correspondence to:

Gary M. Farkas
Department of Psychology
University of Hawaii
2430 Campus Road
Honolulu, Hawaii 96822
Telephone: (808) 948-7644

Running head: Determinants of S[illegible] Error

ADDRESS: Western Psychological Association, 1978

CG012850

Abstract

Several factors thought to influence the representativeness of behavioral assessment data were examined in an analogue study employing a multifactorial design. Systematic and unsystematic methods of observing group behavior were investigated using 18 male and 18 female observers. Additionally, valence properties of the observed behaviors were inspected. Subjects' assessments of a videotape were compared to a criterion code which defined the population of behaviors. Results indicated that systematic observation procedures were more accurate than unsystematic procedures, though this factor interacted with gender of observer and valence of behavior. Additionally, males tended to sample more representatively than females. A third finding indicated that the negatively valenced behavior was overestimated, while the neutral and positively valenced behaviors were accurately assessed.

Observation Procedure, Observer Gender, and Behavior Valence as Determinants of Sampling Error in a Behavior Assessment Analogue

The enterprise of behavior therapy relies largely on the direct observation of behavior, thus the improvement of this "methodology is perhaps the most critical task for strengthening our contribution to the science of human behavior" (Johnson & Bolstad, 1973, p. 8). Underscoring this need, Linehan (Note 1) has called for an examination of the validity of behavioral assessment techniques.

Jones, Reid, and Patterson (1975) have suggested that the definitive features of naturalistic assessment are (a) the sampling of behavior _in situ_, (b) the low degree of inference in describing overt behavior, and (c) the use of trained and impartial observers. However, the _impartiality_ of observers and of the data they record has been recently questioned along three lines: (a) social influence processes, (b) the adequacy of sampling of the population of behaviors, and (c) the gender of observers. In the analogue study to be reported here, these three factors are analyzed, both for their separate effects and their interactions.

Social influence factors have been studied by Scott, Burton, and Yarrow (1967) who found that an observer informed of the nature of the research collected data that confirmed the hypothesis significantly more than did uninformed observers. Because of the importance of this finding to behavioral scientists and because Scott was the informed observer, additional research has been conducted on this bias phenomenon. There is conflicting data as to the effect of social factors on biasing observations (e.g., Kent, O'Leary, Diament, & Dietz, 1974; Shuller & McNamara, 1976; Skindrud, 1973). However, the preponderance of the evidence suggest that under certain conditions both behavioral and global perception data are subject to biasing influence (see also O'Leary, Kent, & Kanowitz, 1975).

An additional variable which may affect the data collection process is the definition of the behavioral code, and the presumed relationship between the code and the personality construct which it is designed to assess. For example, behaviors such as "high-rate" or "self-stimulation" (Jones et al., 1975) are implicitly or explicitly related to constructs such as "hyperactivity," and thus are negatively valenced. Whether observers differentially perceive and record behavior as a function of valence prop[illegible] of code-construct relationships is critical, especially [illegible] most [illegible]ding systems contain behaviors of a range of valences.

A second factor which has been shown to influence the representativeness of behavioral data is the procedure of sampling employed. Thomson, Holmberg, and Baer (1974) have shown that a spaced sequential procedure of observing groups produces the smallest error in estimating the population of behaviors. Powell, Martindale, and Kulp (1975) have demonstrated that momentary time sampling at least once every 120 sec provides data which are truly representative of the population. In general, the more frequent the sampling of the population, the more representative were the data obtained. In a comparison of time-sampling and interval-sampling procedures, Repp, Roberts, Slack, Repp, and Berkler (1976) found that the most representative data were obtained in interval recording, while time-sampling procedures produced highly inaccurate estimates of the population of behaviors. However, both of these latter studies include a number of deficiencies which limit generality. For example, Powell et al. used a simple one-category code of a secretary's in-seat behavior, while Repp et al. employed a time-sampling spacing which yielded infrequent samples of behavior. In addition, because Repp et al. used electro-mechanically-generated chart records as the population from which samples were derived, the artificial nature of this study precludes generalization to research employing human observers: They must decide on the particular person to be sampled as well as the ongoing behavior, and must ensure accuracy of data logging.

Kubany and Sloggett (1973) provided an excellent comparison of the representativeness of two population-sampling procedures. In their study, data collected by a teacher using a VI-4 min schedule of classroom observation were compared to data collected by the experimenter on a FI-15 sec schedule. Kubany and Sloggett's results indicated that momentary time sampling at longer intervals provides data which are veridical with more frequent sampling of the population. A limitation of this study, however, is that the true population of behaviors was not assessed. Thus, there may have been constant error in either over- or underestimation of the sampled behaviors.

The nature of the sampling procedure is particularly important when the behavior of groups is the focus of study (e.g., classrooms). It has long been suspected that teachers are unsystematic in their sampling procedures, and a random manner of observation may lead to inaccuracies in summary reports (Wahler & Leske, 1973) and behavioral data. However, there has been no empirical test of whether systematic methods of sampling improve upon those data collected from unsystematic recording procedures. This question is especially salient because of the recent emphasis on the collection of normative data (e.g., Walker & Hops, 1976).

In addition, one might expect that the valence properties of the behavioral codes employed may interact in some fashion with the observation system used.

A third variable, which has not received attention in the behavioral assessment literature, is the influence of gender on accuracy of observation. Data from the literature on social memory suggest that females make more accurate assessments than males after observing videotapes or photographs of groups of people (Mazanec & McCall, 1975; Witryol & Kaess, 1957). However, these assessments are not made immediately, and thus differ from most behavioral assessment tasks. The vigilance literature suggests that females may be more accurate in tasks requiring sustained attention (Davies & Tune, 1969), although males have been found to have a shorter reaction time in visual vigilance tasks (Harkins, Nowlin, Ramm, & Schroeder, 1974; Tolin & Fisher, 1974) and to be superior in problems which require visual discrimination of objects set in a larger context (Guilford, 1967; Maccoby & Jacklin, 1974a; 1974b). The inconsistency and lack of direct generalizability of the present data base necessitates examination of possible gender differences in the capacity to representatively sample populations of behavior in a naturalistic assessment task.

The present experiment was designed to investigate three factors which possibly influence the representativeness of data obtained in studies employing behavioral assessment in general, and momentary time sampling in particular. This study is not proposed as representative of traditional behavioral data collection; for example, we employed videotapes, rather than live performers, in order to assess the actual population of behaviors. Rather, our purpose is to contribute to knowledge of those factors which may bias observation, through an analogue study with high internal validity, and which allows for higher-order interactions to be revealed. First, bias effects due to valence properties of code-construct relationships were investigated using behavioral codes to represent negative, neutral, and positively valenced personality characteristics. Second, the process of sampling groups was examined, and an unsystematic (idiosyncratic) method of observation was compared with a systematic (standard) procedure. Third, the representativeness of sampling as a function of observers' gender was investigated. Additionally, both algebraic and absolute deviation (Newell, 1976) in sampling the true population of behaviors were assessed. Finally, stepwise regression was used to examine predictors of absolute error.

Method

Subjects

Fifty-three subjects between the ages of 18 and 42 years (mean=22.6) participated in this study as observers. Eighteen male and 18 female observers are included in the final sample; eight males and nine females were excluded on the basis of the selection criteria which are detailed in the Design and Procedure section. Subjects had an average of 2.1 years of college education, with a mean of 3.8 psychology courses. All subjects received course credit for their participation.

Materials

In order to specify the population of behaviors contained within the assessment stimuli, a videotape was produced which depicted six people (two male and four female) who were instructed to follow an informal script. This group tape was 10 min long, and designed to resemble the administration of a group personality questionnaire. The actors sat about a large conference table, and were instructed to behave as if they were taking a test, although they were told to engage in the following behaviors of interest: (a) holding the test booklet in both hands; (b) adjusting clothing, jewelry, and so forth; resting head in both hands; and (c) crossing the arms. These specific behaviors defined the population of behaviors which subjects were to later sample.

After the videotaping was completed, a signal which noted 10 sec intervals was dubbed onto the audio track. A criterion code of all actors' behaviors at all intervals (defined as the instant the signal sounded) was constructed by the author and a female assistant. Disagreements were resolved in consultation with other assistants who were either male or female. The group tape that was employed had relatively equal percentages of intervals during which the actors engaged in the three valenced behaviors (Neutral=10.6%; Negative=8.3%; and Positive=13.1%). Because 68% of the scorable intervals contained nonsalient behaviors, over- or undersampling of one of the behaviors of interest did not necessarily determine the sampling probability of another recordable behavior. All tapes were presented on a black and white video monitor (Sony model CVM-110UA). All subjects recorded their observations on a three-item checklist.

Design and Procedure

This study used a three-way design, with two levels of each between groups factor (observation procedure and observer gender) and three levels of the within subjects factor (code valence). The investigation involved a single laboratory session, with subjects randomly assigned to observation procedure groups. All levels of the within subjects factor were obtained from the single checklist completed by each observer.

Subjects were recruited from two psychology lecture classes, with the experimenter requesting "research assistants" to aid in a behavioral observation project. Upon arriving at the laboratory, subjects completed an "application for assistantship" which requested basic demographic information. After answering preliminary questions, the first instructional tape was presented to the subjects, who were seen in groups numbering up to five.

Experimental deception. A deceptive set was introduced to provide for valenced behavioral codes. Ostensibly, the observers were to code overt behaviors of people taking a group personality test. The study was, supposedly, an attempt to compare two methods of personality assessment, and the observers were told that their observations were to be later correlated with the questionnaire measures. The observers were needed to "ensure that experimental bias would be eliminated, as the experimenter had already scored the personality tests."

Further instructions were delivered on videotape, to insure standardized presentation of information. The behavioral codes were defined and modeled, and a rationale was offered for the scoring of the particular behaviors. It was explained that an earlier study had found relationships between these overt behaviors and the personality constructs which were measured in the test. "Neuroticism," the negatively valenced behavior, ostensibly was related to a category called "Fidget"--the manipulation of clothing or jewelry, or holding the head in both hands. "Positive Functioning," the positively valenced behavior, was said to be related to "Bookholding," or holding the test booklet with both hands. Finally, subjects were told that a "neutral" behavior needed to be observed, that being the "Armcrossing" category which was supposedly related to no particular personality construct.

Training tapes. To ensure that all subjects could accurately assess the behaviors of interest, three videotapes of individual females were used for training purposes. During each of these 5 min tapes, questions could be asked, and definitions of the behaviors were again offered. A fourth videotape was employed to assess accuracy of code usage, with an interval-by-interval agreement score (Hawkins & Dotson, 1975) of 90% required in order to view the group tape.

Observation procedure and group tape. After the subjects' accuracy was assessed, a second instructional videotape was presented which explained the coding procedure to be used in assessing a group of individuals. Half of the observers were taught to observe the group systematically. ("Start with any individual, then observe each person for one interval in a clockwise fashion until the tape ends.") The remaining observers were asked to use an unsystematic observation strategy. ("Start with any individual, choose the next person to code in a completely random manner; do not use any system which would allow prediction of the next person you will choose.") All subjects were then cautioned that they must observe the individuals in the group in systematic/unsystematic fashion "to ensure that the data will be free from bias."

Immediately before observing the group tape, all subjects were asked to define the behaviors of interest, and to state the observation procedure to be employed. Each subject observed the group tape in isolation; all were told that they were viewing a unique tape.

After the group tape was completed, a questionnaire was given to check the adequacy of the deceptive set. Subjects were asked first to state the relationships between the behaviors and the personality constructs, and second, to state the degree of their beliefs in these relationships. Only those subjects who stated at least two of the relationships and indicated at least a mild belief in them were included in the present analysis.

Results

Of initial interest was the assessment of possible differences between subjects retained or excluded from the primary analyses on the basis of rejection criteria. The *t* tests performed indicated no group differences with respect to age, class standing, number of psychology courses taken, grade-point average (GPA), or pretest accuracy. The largest (though still nonsignificant) difference was found on the absolute error variable, with subjects included in the analysis making an average absolute error of 10.7% while excluded subjects deviated 8.9%, $t(51) = -1.66$, $p > .10$. Deviation for this and all other analyses was defined as the observers' estimation of the percentage of intervals in which the group engaged in specific behavior minus the true percentage of intervals the group engaged in these behaviors. In addition, if compared to the actual population of behaviors, a relative indication of over- and underestimation is available.

Prior to analyses of the primary dependent measures, a two-way analysis of variance (ANOVA) was used to determine whether each observation/gender group retained in the analyses were equated on demographic, measures of deception, and accuracy of pretest variables. There were no significant group differences on age, class standing, GPA, pretest accuracy, number of relationships remembered, or the degree of belief in these relationships. For all subjects, the average belief rating of the negatively valenced category was 2.94, with the positively valenced category rated 2.77 (1=not related, 5=very strongly related), $t(70)=.87$, $p>.10$. However, one ANOVA did reveal a significant source, with females having taken more psychology courses than males, $F(1,32)=5.66$, $p<.025$.

When analyzing deviation scores, the dependent measure may be conceptualized as either the long-range average of observers' error (algebraic error) or the average error independent of direction (absolute error). In the present experiment all analyses employing algebraic error were analyzed via three-way ANOVA. Absolute error was considered primarily relevant to the between group factors, and was analyzed using a two-way ANOVA.

Analysis of algebraic error indicated several significant sources of variance. Systematic observers were more accurate than unsystematic observers, with the former making deviation errors of -.23%, and the latter making errors of 2.57%, $F(1,32)=16.09$, $p<.001$. Thus, in terms of total recordable behaviors, unsystematic observers perceived 3.7% more behaviors than were actually present, while systematic observers actually underestimated the total behavior by .3%. The observation procedure, however, interacted with the observers' gender, $F(1,32)=4.28$, $p<.05$ (see Figure 1), though it did not interact with code valence. Category valence was a third source of significance, $F(2,64)=13.92$, $p<.001$, with an a posteriori t test indicating that more deviation errors were made in observing the negatively valenced category (3.52%) than either the positively (-.18%) or neutrally (.17%) valenced categories, $t(64)=4.98$, $p<.001$. In terms of the population of behaviors, observers overestimated the negatively valenced and neutral category by 42.4% and 1.6%, respectively, while underestimating the positive category by 1.3%. Additionally, a valence x observation procedure x gender interaction was significant, $F(2,64)=5.03$, $p<.01$ (see Figure 2).

Insert Figure Figure 2 approximately here

The two-way ANOVA for absolute error indicated that males were more accurate in sampling the population than females, $F(1,32)=6.95$, $p<.015$. Males deviated an average of 9.06% while females deviated an average of 12.28%. There were no additional sources of variance for absolute error.

Stepwise regression analyses were performed on both absolute error, as well as the number of errors made on the pretest (accuracy of using the codes). After controlling for the sources of error which constituted the composite absolute error score, no demographic variables predicted absolute error. However, GPA was a significant predictor of errors on the pretest, with a partial correlation of -.39, $F(1,34)=6.17$, $p < .05$. Thus, the lower the GPA, the more errors were made on the test tape. Of additional importance, males tended to have lower grade point averages, although this was not a significant source of variance ($p > .10$).

Discussion

It appears that systematic momentary time-sampling procedures provide relatively unbiased information, while unsystematic time sampling produces a significantly larger degree of population-sampling error. This result extends the work of Powell et al. (1975), and partially confirms speculation by Omark, Fiedler, and Marvin (1976) that unsystematic methods of observation allow idiosyncratic bias. The degree this idiosyncratic bias is evidenced by a significant gender x valence x observation procedure interaction, and inspection of Figure 2 suggests that sampling error is actually multidetermined. However, no interpretation of this interaction will be attempted. We would like to note, though, that behavioral assessment, like other subdisciplines of psychology, appears to be subject to higher-order interactions (cf., Cronbach, 1957; 1975) and continued investigation of attributes and treatments which may effect observer behavior appears to be in order.

An unexpected finding was that females tended to observe less representatively than males when absolute error is the measure of interest. When algebraic error is inspected, females appear to be less representative observers when using a random method of observation. The lack of main effect for observer gender in the algebraic error analysis reflects the dispersion of females' error in both the positive and negative direc[illegible] pr[illegible] a level of error approaching zero. [illegible]tionale for these sex differences may be found in the literature on visual discrimination (Maccoby & Jacklin, 1974a; 1974b). However, the present findings should be considered only as suggestive for further research, and certainly not as an adequate basis for personnel decisions.

An additional note of importance is the failure to detect differences in the rate of absolute error between subjects included or excluded from the primary analyses. While this suggests a degree of generality of our findings, a more sensitive test would have required a larger sample of subjects who did not believe in the deceptive code-construct relationships.

In conclusion, valence of behavioral codes, gender of observers, and the observation procedure have been demonstrated to influence the representativeness of momentary time sampling in a behavioral assessment analogue. It is clear that if behavior therapists are to continue to rely on their data as evidence of the efficacy of their techniques, rather than on testimonials (Johnson & Bolstad, 1973), they must eliminate or control nonrandom sources of population-sampling error. First, however, the complexities inherent in behavioral observation must be thoroughly studied.

It is likely that the effects of social influence and gender can be negated by careful observer training, just as the adoption of systematic sampling techniques has produced more valid estimates of time frequencies. The value of analogue studies lies in identification of those biasing factors to which trainers should attend.

Reference Notes

1. Linehan, M. M. *Content validity in behavioral assessment.* Paper presented at the 84th annual meeting of the American Psychological Association, Washington, D. C., September 3, 1976.

References

Cronbach, L. J. Beyond the two disciplines of scientific psychology. American Psychologist, 1975, 30, 116-127.

Cronbach, L. J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Davies, D. R., & Tune, G. S. Human vigilance performance. New York: American Elsevier, 1969.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill, 1967.

Harkins, S. W., Nowlin, J. B., Ramm, D., & Schroeder, S. Effects of age, sex, and time-on-watch on a brief continuous performance task. In E. Palmore (Ed.), Normal Aging (Vol. II). Durham, North Carolina: Duke University Press, 1974.

Hawkins, R. P., & Dotson, V. A. Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of inter-observer agreement scores in interval recording. In E. Ramp & G. Semb (Eds.), Behavior analysis: Areas of research and application. Englewood Cliffs, N. J.: Prentice-Hall, 1975.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts and practice. Champaign, Illinois: Research Press, 1973.

Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. III). San Francisco: Jossey-Bass, 1975.

Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. Expectation biases in observational evaluation of therapeutic change. Journal of Consulting and Clinical Psychology, 1974, 42, 774-780.

Kubany, E. S., & Sloggett, B. B. Coding procedure for teachers. Journal of Applied Behavior Analysis, 1973, 6, 339-344.

Maccoby, E. E., & Jacklin, C. N. The psychology of sex differences. Stanford: Stanford University Press, 1974. (a)

Maccoby, E. E., & Jacklin, C. N. Myth, reality, and shades of gray: What we know and don't know about sex differences. Psychology Today, 1974, 8(7), 109-112. (b)

Mazanec, N., & McCall, G. J. Sex, cognitive categories, and observational accuracy. Psychological Reports, 1975, 37, 987-990.

Newell, K. M. More on absolute error, etc. Journal of Motor Behavior, 1976, 8, 139-142.

O'Leary, K. D., Kent, R. N., & Kanowitz, J. Shaping data collection congruent with experimental hypotheses. Journal of Applied Behavior Analysis, 1975, 8, 43-51.

Omark, D. R., Fiedler, M. L., & Marvin, R. S. Dominance hierarchies: Observational techniques applied to the study of children at play. Instructional Science, 1976, 5, 403-423.

Powell, J., Martindale, A., & Kulp, S. An evaluation of time-sampling measures of behavior. Journal of Applied Behavioral Analysis, 1975, 8, 463-469.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. A comparison of frequency, interval, and time-sampling methods of data collection. Journal of Applied Behavior Analysis, 1976, 9, 501-508.

Scott, P. M., Burton, R. V., & Yarrow, M. R. Social reinforcement under natural conditions. Child Development, 1967, 38, 53-63.

Shuller, D., & McNamara, J. R. Expectancy factors in behavioral observation. Behavior Therapy, 1976, 7, 519-527.

Skindrud, K. Field evaluation of observer bias under overt and covert monitoring. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, Illinois: Research Press, 1973.

Thomson, C., Holmberg, M., & Baer, D. M. A brief report on a comparison of time-sampling procedures. Journal of Applied Behavior Analysis, 1974, 7, 623-626.

Tolin, P., & Fisher, P. G. Sex differences and effects of irrelevant auditory stimulation on performance of a visual vigilance task. Perceptual and Motor Skills, 1974, 39, 1255-1262.

Wahler, R. G., & Leske, G. Accurate and inaccurate observer summary reports. Journal of Nervous and Mental Disease, 1973, 156, 386-394.

Walker, H. M., & Hops, H. Use of normative peer data as a standard for evaluating classroom treatment effects. Journal of Applied Behavior Analysis, 1976, 9, 159-168.

Witryol, S. L., & Kaess, W. A. Sex differences in social memory tasks. Journal of Abnormal and Social Psychology, 1957, 54, 343-346.

Footnotes

This study was truly the result of the effort of a number of individuals, and the author is indebted to the entire staff of the Kamehameha Early Education Program for facilitating its outcome. In particular, Larry Loganbill was an outstanding audio-visual resource, Lou Ann Wooddell showed dedicated acting and coding ability, and Mary Hinck deserves plaudits for her consistent and good-humored help. Kim C. M. Sloat was instrumental in all phases of this project, and his support is gratefully acknowledged.

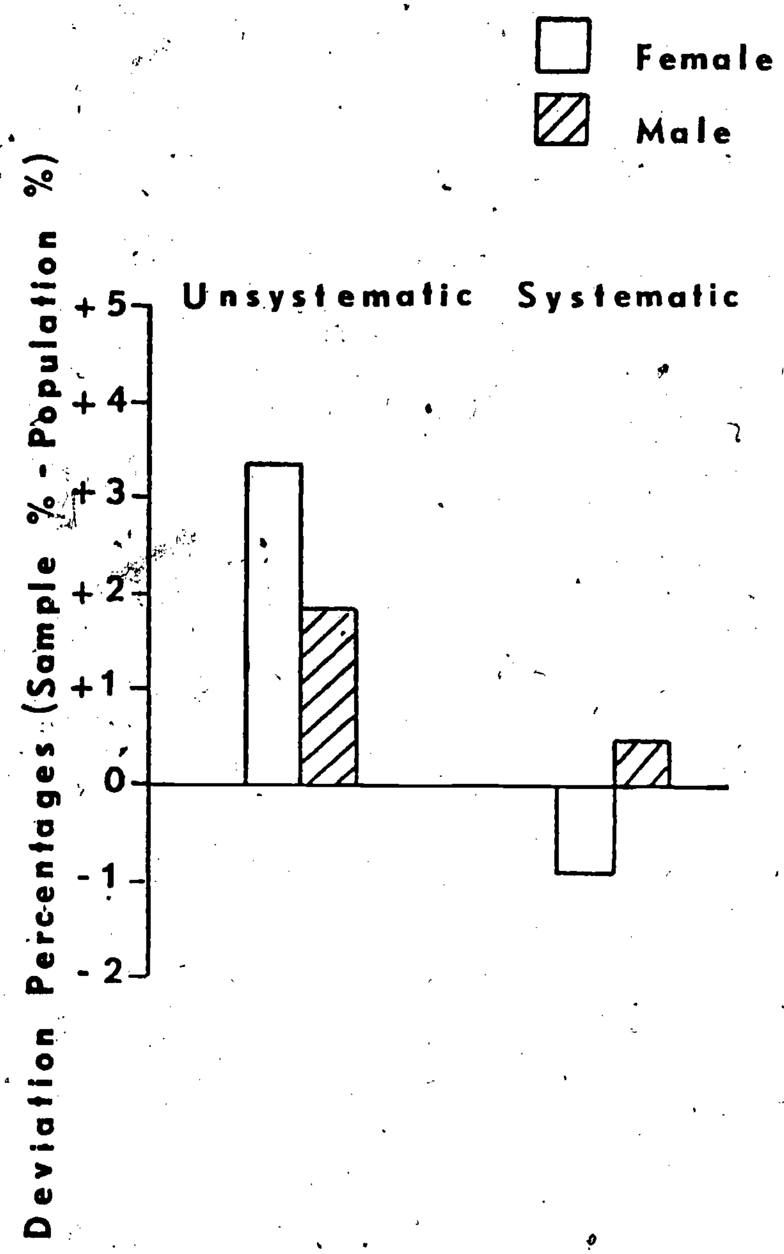
This paper is part of a thesis submitted by the first author to the graduate division of the University of Hawaii in partial fulfillment of the requirements for the M.A. degree in psychology.

Requests for reprints of this article may be addressed to Gary M. Farkas, Department of Psychology, University of Hawaii, 2430 Campus Road, Honolulu, HI, 96822.

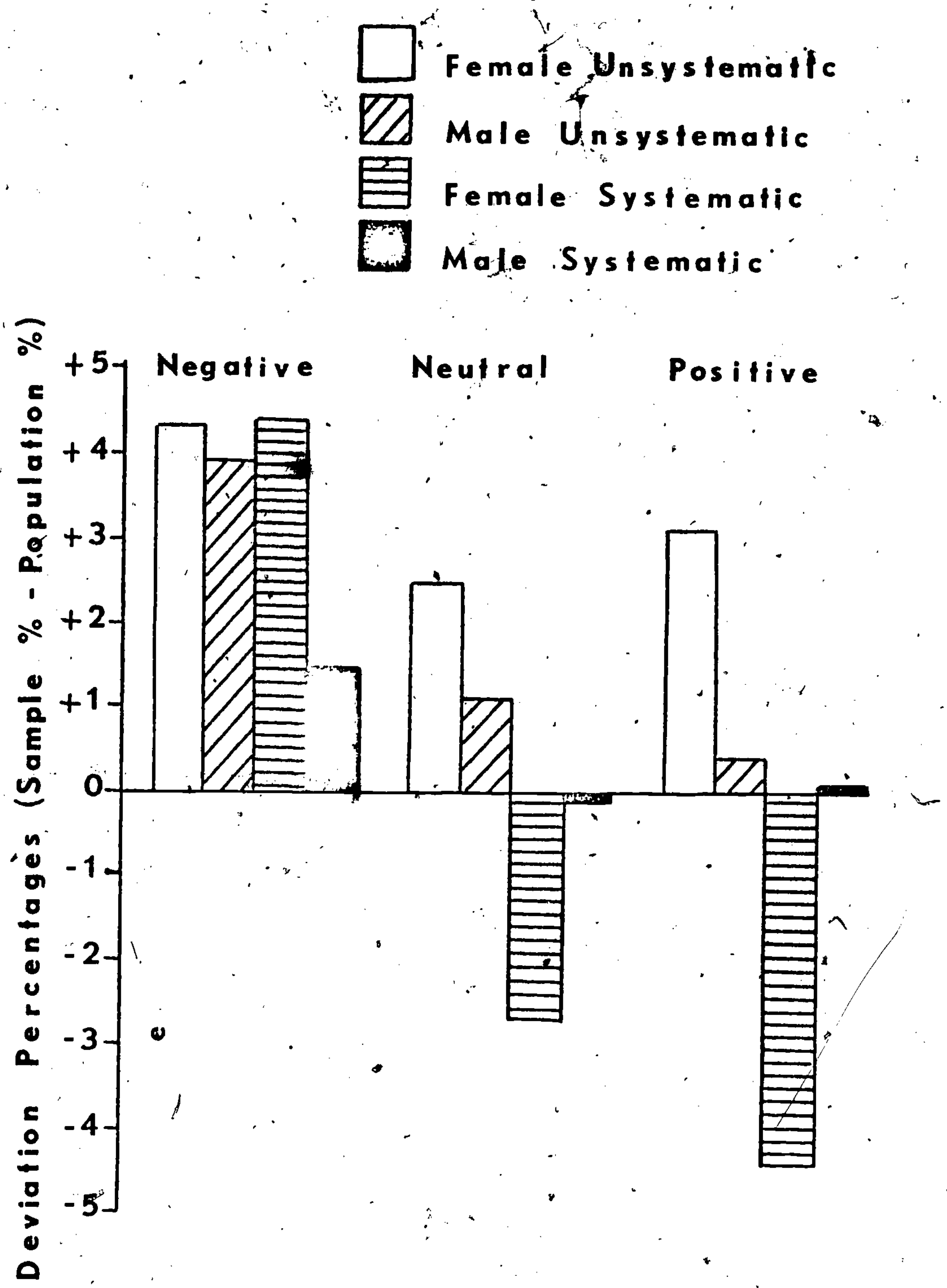
Figure Captions

Figure 1. Sampling Error as a Function of Gender and Observation Procedure.

Figure 2. Sampling Error as a Function of Gender, Observation Procedure, and Valence.



Female
Male
Unsystematic
Systematic
Deviation Percentages (Sample % - Population %)
+5
+4
+3
+2
+1
0
-1
-2


Female Unsystematic
Male Unsystematic
Female Systematic
Male Systematic
Deviation Percentages (Sample % - Population %)
Negative
Neutral
Positive
+5
+4
+3
+2
+1
0
-1
-2
-3
-4
-5