DOCUMENT RESUME

ED 102 205 TM 004 264

AUTHOR Hollenbeck, Albert R.; Slaby, Ronald G.
TITLE Measures of Reliability in Behavioral Observation: The Advantage of "Real Time" Data Acquisition.
SPONS AGENCY Washington Univ., Seattle. Child Development and Mental Retardation Center.
PUB DATE [74]
NOTE 21p.
AVAILABLE FROM Albert R. Hollenbeck, Department of Psychology, N.I-25, University of Washington, Seattle, Washington 98195

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Codification; *Digital Computers; Infant Behavior; Measurement Techniques; *Observation; Test Reliability; Video Tape Recordings
IDENTIFIERS Behavioral Observation Scoring System

ABSTRACT

Two observers who were using an electronic digital data acquisition system were spot checked for reliability at random times over a four month period. Between-and within-observer reliability was assessed for frequency, duration, and duration-per-event measures of four infant behaviors. The results confirmed the problem of observer drift--the fluctuations of scores across sessions--for the frequency and duration-per-event measures. In contrast, the "real time" duration scores were stable across sessions, indicating the robustness of this measure of behavior. (Author)

MEASURES OF RELIABILITY IN BEHAVIORAL OBSERVATION: THE ADVANTAGE OF "REAL TIME" DATA ACQUISITION

Albert R. Hollenbeck and Ronald G. Slaby
University of Washington

Abstract

Two observers who were using an electronic digital data acquisition system were spot checked for reliability at random times over a four month period. Between- and within-observer reliability was assessed for frequency, duration, and duration-per-event measures of four infant behaviors. The results confirmed the problem of observer drift--the fluctuations of scores across sessions--for the frequency and duration-per-event measures. In contrast, the "real time" duration scores were stable across sessions, indicating the robustness of this measure of behavior.

MEASURES OF RELIABILITY IN BEHAVIORAL OBSERVATION: THE ADVANTAGE OF "REAL TIME" DATA ACQUISITION[1]

Albert R. Hollenbeck[2] and Ronald G. Slaby[3]
University of Washington

A number of recent studies have been addressed to the problem of assessing reliability in observational research (e.g., Reid, 1970; Johnson and Bolstad, 1973; Whelan, 1974). This interest in the more subtle aspects of the observational process reflects a general increase in the use of direct observational measures in psychological research. Reliability in observational research generally means that two or more observers independently record the same naturally occurring behavioral events in a similar way. It has been generally assumed that if two observers achieve a high level of inter-observer reliability, then they are measuring the same aspects of behavior across sessions as well. Yet, Reid (1970) has demonstrated that this may not be the case. His data indicate that whereas reliability between observers can remain at a constant level, reliability across sessions for a given observer tends to decrease with time. This problem of observer drift, i.e., the fluctuation of observations across sessions, has important implications both for the interpretations of data already collected and for the collection of data in future research.

A second problem in establishing reliability is that of selecting an appropriate statistical index. The typical measure of reliability has been some summary statistic such as a percentage agreement score or a correlation coefficient. It has been generally assumed that these traditional statistical measures used to compute reliability are valid. However, percentage agreement and correlation measures of reliability have recently come under justifiable criticism (e.g., Hartmann, 1974) for over-estimating reliability as well as for being insensitive to the detection of chance agreements.

In order to overcome these difficulties in establishing reliability, systematic attempts have been made to evaluate different components of the observational process (e.g., Mash, 1973; Taplin and Reid, 1973; Hawn, Brown, and LeBlanc, 1973). These evaluations have uncovered several questions underlying the basic assumptions of the behavioral observation methods. For example, in the majority of observational studies, the most common measure used has been the frequency of occurrence of some behavioral event. In fact, most studies have been time-sampled in such a way that actual frequencies are not scored. Rather, a modified-frequency score (i.e., a score based on the number of arbitrarily defined time intervals in which an event has occurred) is used to mark simple occurrence or non-occurrence of a behavioral event. The very nature of modified-frequency measurement is suspect, since actual frequencies and durations are confounded.

Recent advances in technology have provided electronic systems which allow the unconfounded recording of actual frequency and duration scores. MIDCARS and the Behavioral Observation Scoring System (BOSS) are two such systems (Sackett, Stephenson, and Ruppenthal, 1973). These advances, which allow the experimenter to measure exact frequencies and durations separately, raise several interesting questions. How does the reliability of real frequency and real duration measures compare with that of the modified-frequency measures typically used in previous research? Are the unconfounded frequency and duration measures subject to fluctuations in observations across sessions, as reported by Reid and others for modified-frequency scores? The purpose of this study was to examine the reliability of observations based on real frequency and real duration measures. This study was designed to assess observer drift in the reliability of these measures.

Method

Subjects

Two female undergraduates at the University of Washington served as observers in this study. Both observers were volunteers who received academic credit for applied field work in psychology.

Apparatus

A videotape machine was used to record the responses of a six-month-old infant to a stimulus presentation designed to elicit vocalizations from the infant. This videotaped sequence was presented to the observers for purposes of assessing observer reliability.

The behavior code was a modified version of the one previously used by Hollenbeck (1971). Five mutually exclusive and exhaustive behavioral categories--Vocalization, Head Movement, Arm Movement, Body Movement, and No Behavior--were hierarchically arranged and scored on a priority basis. Specifically, Vocalizations made by the infant took scoring preference over Head Movements when both behaviors occurred simultaneously. In the same way, Head Movements were scored over Arm Movements; Arm Movements over Body Movements; and any movements or vocalization took scoring preference over the No Behavior category.

The Behavioral Observation Scoring System (BOSS) was used to record the coded data. BOSS is an electronic digital data acquisition system developed at the University of Washington Child Development and Mental Retardation Center and the University of Washington Primate Center. This system allows behavioral events to be recorded in terms of their actual frequencies and durations and stored electronically on a magnetic cassette audio tape recorder. The cassette data tapes can then be played through an appropriate interface into a computer for analysis of the data. A detailed description of BOSS is presented in Sackett, et al. (1973).

Procedure

The observers were recruited from an undergraduate psychology course by means of an announcement asking for volunteers to participate in an observational study of infants. Academic credit in independent field work was offered at a later time.

Training. The observers were trained in four phases. First, Observer A coded the videotape sequence stating each code aloud as it occurred. Observer B then attempted to follow Observer A's coding, but using her own choice of codes where disagreements occurred. Second, the two observers discussed their disagreements with the experimenter after each coding session. All disagreements were resolved by mutual agreement. Third, the procedure was reversed and Observer B stated the code while Observer A recorded silently. Finally, a third pass through the videotape was made with each observer recording silently and independently. This entire training procedure was repeated twice a week for one month. At the end of the training period observers were presented a new segment of the videotape and asked to code the tape independently. On two successive codings of new material the observers achieved frequency percentage agreement scores greater than 90 per cent for each trial. The first two checks after criterion agreement was reached consisted of part-new and part-old segments of the videotape. The mean percentage agreement between the two observers for the frequency scores of the five behavioral categories was 97.8 per cent and 94 per cent, respectively. These percentages were significantly greater than a pre-established criterion of 80 per cent agreement. Duration measures of reliability were not computed.

Data collection. Each observer was instructed that the primary purpose of the study was to gather information about infants. Observers were told that at random intervals their observer agreement would be checked; however, they received no advance warning of the checks. During the four months after the initial training the observers were "spot checked" five times for reliability on the same segment of the videotape sequence. Spot checking is a commonly employed procedure whereby reliability is assessed periodically rather than continuously (see Taplin and Reid, 1973). In this case the duration between the five checks varied from two to four weeks. The same segment of videotape was used for each check and all checks were taken independently for each observer. Between sessions, observers actually scored the behavior of infants participating in the infant research project. This procedure, with its long and variable duration between checks and its interposed coding task, was designed to minimize observer expectation and simple recall. In fact, the observers verbally reported a vague sense of what was on the videotape, but had a difficult time recalling any specifics.

Data analysis. Measures of (1) frequency, (2) duratinn-per-event, and (3) duration were taken from the same presentation for a standard trial length (5.5 minutes). A multiple regression analysis using backward deletions (in which each main and interaction factor was sequentially deleted from the total variance) was performed on each of the three dependent measures, as suggested by Cohen (1968). The variables in these regression analyses included Observers (2), Sessions (5), 4 separate behavioral categories, and their interactions. In addition, a trend analysis across sessions was included. Based on these regression analyses three analyses of variance were computed.

Results

Frequency data. The analysis of variance for frequency scores is presented in Table 1. The analysis revealed a significant linear trend ($p < .001$) across sessions. Further variation across sessions was characterized by a significant quartic trend ($p < .025$). Each of the four behavioral categories (Vocalizations, Head Movements, Arm Movements, and Body Movements) differed from the category of No Behavior against which they were contrasted. Finally, the Observer X Vocalization interaction was significant ($p < .001$), indicating variation between observers in their recording of frequencies of Vocalization in contrast to those of the No Behavior category. Observer B scored Vocalization more frequently than Observer A. The regression analysis for frequency scores revealed that a significant amount of variability ($R^2 = .95$) was accounted for by the four behavioral categories tested against the category of No Behavior.

Insert Table 1 about here.

Duration-per-event data. The analysis of variance for duration-per-event scores is presented in Table 2. The findings for duration-per-event scores were similar to those for the freque.icy scores.

Insert Table 2 about here.

Specifically, a significant linear trend ($p < .001$) was revealed, indicating significant variation across sessions. Further variation across sessions was characterized by a significant quadratic trend ($p < .05$). Each of the four behavioral categories differed from the category of No Behavior against which they were contrasted. Observers showed significant overall differences ($p < .001$) in their durations-per-event scores across all behavioral categories and all sessions. Observer A scored longer durations-per-event than Observer B. In addition, the Observer X Vocalization and the Observer X Body Movements interactions were significant, indicating variation between observers in their recording of the durations-per-event of these two behaviors in contrast to those of the No Behavior category. As was the case for frequency scores, a significant amount of variability ($R^2 = .83$) was accounted for by the four behavioral categories tested against the category of No Behavior.

Duration data. The analysis of variance for duration scores is presented in Table 3. In contrast to both the frequency and the duration-per-event measures, the duration measure was stable across sessions.

Insert Table 3 about here.

Again, each of the four behavior categories differed from the category of

No Behavior against which they were contrasted. Although observers showed no overall differences in their duration scores, Observer X Behavior Category interactions were significant for each of the four behaviors in contrast to the category of No Behavior. The regression analysis for duration scores revealed that a significant amount of variability ($R^2 = .77$) was accounted for by the four behavioral categories tested against the category of No Behavior.

Discussion

These results confirm previous findings of observer drift, i.e., the fluctuation of observations across sessions. Consistent with Reid's (1970) finding of observer drift for modified-frequency scores, the unconfounded real frequency score used in the present study showed large fluctuations across sessions (see Figure 1). In addition, considerable observer drift was noted for the duration-per-event measure. The pattern of fluctuation of these scores was not characterized by a sharp decrement followed by a stable level of performance, as found by Taplin and Reid (1973). Rather, these scores showed intermixed rises and declines across sessions, as indicated by significant quartic and quadratic trends for frequency and duration-per-event measures, respectively. One possible explanation for this additional fluctuation is that the amount of time between sessions was both longer and more variable than has been characteristic of previous studies of reliability.

In contrast to the findings for the dependent measures directly related to frequency data (i.e., modified-frequency, real frequency, and duration-per-event measures), real duration scores showed no observer drift (see Figure 1). Duration scores were generally stable across sessions. The greater stability of duration scores may be attributable to several factors. It may be that across sessions observers tend to discriminate an increased number of discrete events, each of shorter duration. This is suggested in the present findings by an increase across sessions in frequency scores and a concurrent decrease across sessions in the duration-per-event scores. However, provided that event frequencies are recorded in basically the same categories over sessions, total duration scores for each category would be expected to remain relatively unaffected by this trend toward finer discrimination events.

A second factor contributing to greater stability of duration scores is that duration, unlike frequency, is by definition a weighted measure. Specifically, a duration score is more heavily weighted than a frequency score to the extent that the durations of observable events are long. The longer the durations-per-event are for a given behavior, the heavier is the weighting of the duration score as compared to the frequency score. Thus, minor fluctuations in scoring across sessions would be expected to affect duration scores (with their greater weight) relatively less than frequency scores. Figure 1 illustrates the finding that the percentage difference of session means from the grand mean is relatively smaller for duration scores than for frequency scores.

Insert Figure 1 about here.

An interesting secondary finding suggests alternative means of assessing between-observer reliability. It was found that the majority of the variance was accounted for among the Behavioral Categories being coded, rather than between Observers. For frequency, duration-per-event, and duration scores, respectively, the proportion of the variance accounted for among behavioral categories was .97, .83, and .77; whereas the proportion of the variance accounted for between Observers was .02, .11, and .002. This implies that the between-observer reliability was high for all three measures. Traditional measures of reliability support this notion insofar as both the average percentage agreement and the average correlation coefficient between observers was greater than .90 for frequency scores obtained in the first training session. Nevertheless, the analyses of variance revealed that observers differed significantly in their recording of at least one behavior for each of the dependent measures. Furthermore, observers showed overall differences across all four behaviors in their duration-per-event scores. These findings indicate that the analyses of variance provide a more sensitive test of differences between observers than do the traditional measures of between-observer reliability.

One possible point of criticism specific to this analysis is that total duration summed across all coding categories is completely determined by the standard trial length. Since total duration cannot vary from session to session, one might conclude that the reported stability of duration across sessions is trivial. However, the reported stability was based on four behavior codes which together accounted for an average of only 40 per cent of the total duration; the category of No Behavior accounted for the other 60 per cent. Since the duration scores for the four behavior categories were thus free to vary, the stability of duration of these individual behaviors across sessions was a meaningful finding.

Taken together, these findings suggest that to properly interpret measures of reliability the robustness of "real time" duration measures must be considered. Based on the findings of the present study, duration measures appear to be less suceptible to observer drift. In addition, assessment of reliability through analyses of variance and regression analyses should be further explored, considering the potential advantages in precision and sensitivity.

References

Cohen, Jacob. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Hartmann, Donald P. Assessing the quality of observational data. Paper presented at the Meeting of the Western Psychological Association, San Francisco, April, 1974.

Hawn, Joyce, Brown, George and LeBlanc, Judith M. A comparison of three observation procedures: consecutive intervals on-the-spot; consecutive intervals from video tape; 10-sec-on, 10-sec-off from video tape. Paper presented at the Meeting of the American Psychclogical Association, Montreal, August, 1973.

Hollenbeck, A. R. Imitation as a skill in infancy. Unpublished manuscript, University of Washington, Seattle, 1972.

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

Mash, Eric J., and McElwee, John D. Situational effects of observer accuracy: Behavioral predictability, prior experience, and complexitv of coding categories. Child Development, in press.

Reid, John B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Sackett, G. P., Stephenson, E., and Ruppenthal, G. C. Digital data acquisition systems for observing behavior in laboratory and field settings. Behavioral Research Methodology and Instrumentation, 1973, 5, 344-348.

Taplin, P. S., and Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. Child Development, 1973, 44, 547-554.

Whelen, B. Reliability of Human Observers. Unpublished doctoral dissertation, University of Utah, Salt Lake City, Utah, June, 1974.

Footnotes

[1] An earlier version of this paper was presented at the 82nd Annual Meeting of the American Psychological Association, Division 25, August 30-September 3, 1974, New Orleans, Louisiana. This research was supported, in part, by the Child Development and Mental Retardation Center of the University of Washington.

[2] Requests for reprints should be sent to Albert R. Hollenbeck, Department of Psychology, NI-25, University of Washington, Seattle, Washington 98195.

[3] The authors wish to express appreciation to Gene Sackett for his guidance throughout all phases of this research, and to Beverly Davis and Lynn Davis for their work as observers in this study.

Table 1

Analysis of Variance for Frequency

Source	df	MS	F
Total	12	1420.51	
Residual (error)	37	5.04	
Observer (O)	1	7.22	ns
Vocalizations (V)	1	8217.62	1530.48***
Head Movement (H)	1	28.03	5.56**
Arm Movement (A)	1	1520.07	301.64**
Body Movement (B)	1	6993.80	1387.66**
Trends			
Linear Trend (T)	1	153.76	30.51**
Quartic Trend	1	28.28	5.61**
Cubic Trend[a]			
Quadratic Trend[a]			
O X T	1	12.96	ns
O X V	1	36.98	7.34***
O X H	1	20.83	4.13*
O X A	1	6.67	ns
O X B	1	20.00	3.97*

[a] The quadratic and cubic trends were eliminated from the analysis by t he computer program due to the small MS attributable to these factors.

*p < .05

**p < .025

***p < .001

Table 2

Analysis of Variance for
Duration-per-Event

Source	dr	MS	F
Total	14	611.41	
Residual (error)	35	5.56	
Observer (O)	1	121.68	21.88***
Vocalizations (V)	1	1335.28	240.16***
Head Movement (H)	1	963.33	173.26***
Arm Movement (A)	1	3168.27	569.83***
Body Movement (B)	1	2690.90	483.97***
Trends			
Linear Trend (T)	1	156.25	28.10***
Quadratic Trend	1	30.18	5.43*
Cubic Trend	1	1.00	ns
Quartic Trend	1	16.05	ns
O X T	1	13.69	ns
O X V	1	30.42	5.47*
O X H	1	.83	ns
O X A	1	1.67	ns
O X B	1	33.80	6.08**

*p < .05

**p < .025

***p < .001

Table 3

Analysis of Variance for Duration

Source	df	MS	F
Total	14	1761226.24	
Residual (error)	35	2410.72	
Observer (O)	1	67.28	ns
Vocalizations (V)	1	11981.52	4.97*
Head Movement (H)	1	1061824.50	520.06***
Arm Movement (A)	1	5759801.70	2389.26***
Body Movement (B)	1	17611891.00	7333.03***
Trends			
Linear (T)	1	27.00	ns
Quadratic	1	96.00	ns
Cubic	1	7.00	ns
Quartic	1	4.00	ns
O X T	1	27.00	ns
O X V	1	12200.00	5.06*
O X H	1	49045.00	20.34***
O X A	1	23602.00	9.79**
O X B	1	61829.00	25.65***

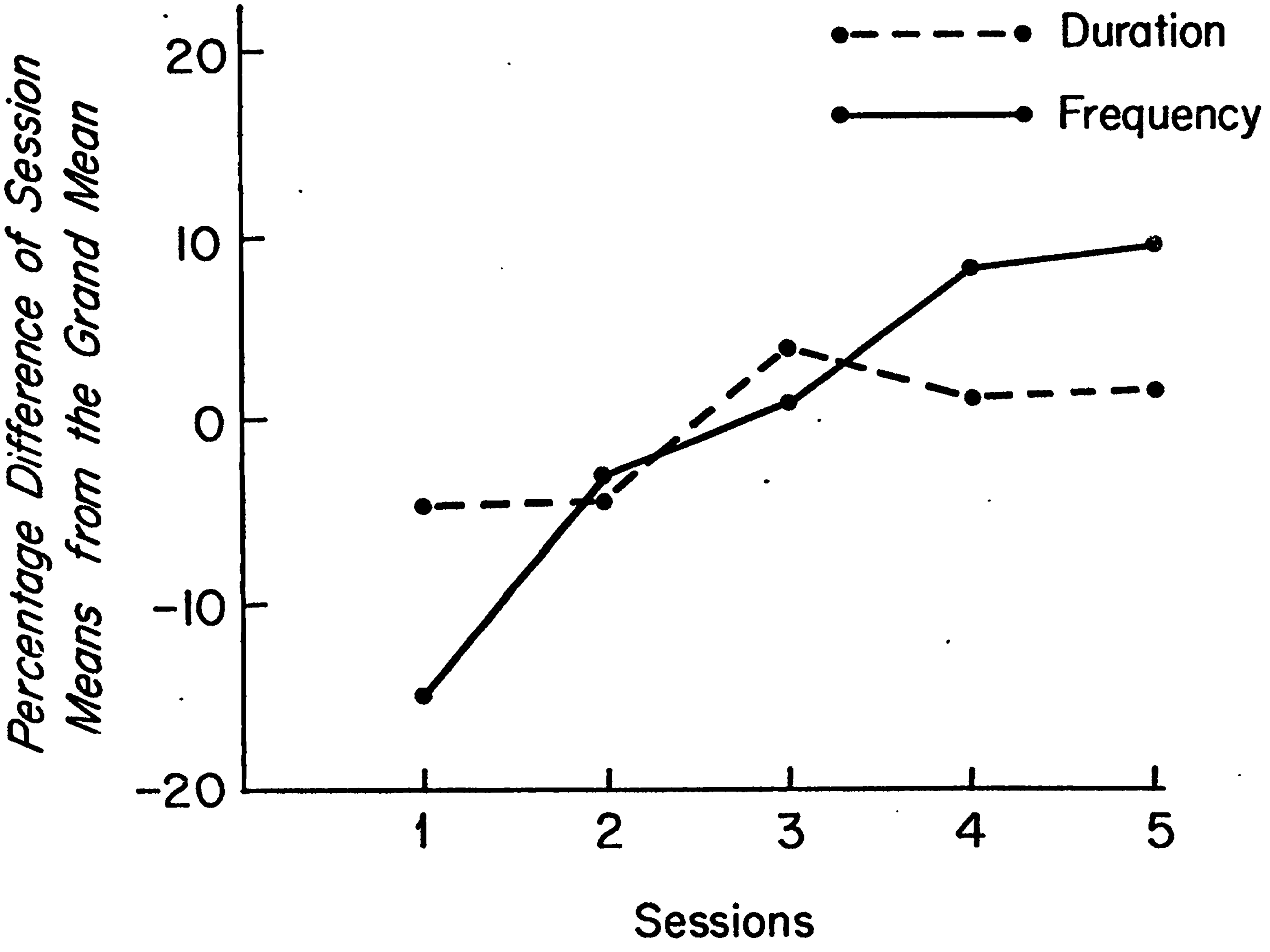
*p < .05

**p < .01

***p < .001

Figure Caption

Fig. 1. Percentage diffeience o: session means (based on the four infant behaviors) from the grand mean for frequency and duration scores.


Duration
Frequency
Percentage Difference of Session Means from the Grand Mean
20
10
0
-10
-20
1
2
3
4
5
Sessions