2003

Computerized Behavioral Observation: Implications for Improving Accuracy and Reducing Bias

Charles Sicotte
University of Rhode Island

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

---

Recommended Citation

Sicotte, Charles, "Computerized Behavioral Observation: Implications for Improving Accuracy and Reducing Bias" (2003). *Open Access Dissertations*. Paper 946.
https://digitalcommons.uri.edu/oa_diss/946

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

COMPUTERIZED BEHAVIORAL OBSERVATION: IMPLICATIONS FOR
IMPROVING ACCURACY AND REDUCING BIAS

BY

CHARLES SICOTTE, M.A.

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2003

Abstract

Direct observation of behavior is a form of behavioral assessment that maintains a low level of inference when applied to psychoeducational decision-making. The recent proliferation of computer behavior-observation systems has provided observers with an efficient means for collecting direct behavior observational data (Kahng & Iwata, 2000; Shapiro & Kratochwill, 2000a, 2000b). Although computerized methods are less cumbersome than other recording methods (e.g., partial-interval form, narrative recording), it is not clear if computerized methods offer improvement by enhancing observational accuracy or reducing judgment bias that is generally initiated in applied settings (e.g. referral information; Arkes & Harkness, 1980; O'Reilly, Northcraft, & Sabers, 1989). Undergraduate participants ($N = 243$) were trained to perform a 10-minute observation using a computer observation program, partial interval form, narrative recording method, or a no-recording method. In addition, the participants were randomly given an internalizing, externalizing, or a non-specific psychopathology referral. The participants who used computerized and interval-recording methods consistently produced more accurate observations and

were less susceptible to referral bias than the narrative and no-method groups. Implications for applied practice and suggestions for future research are discussed.

Acknowledgements

There are many people that helped make this research possible that I would like to acknowledge. I would like to thank my brother, Brad Sicotte, for offering his expertise in constructing the computer observation program for this study. Dr. W. Grant Willis and Dr. David Faust provided valuable guidance and feedback for this work. In addition, Kathryn Knuettel and Tom Kelley graciously volunteered to helping me score and enter data. Finally, I would like to thank my wife Michele and my daughter Danielle for their love and encouragement.

Table of Contents

Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	viii
List of Figures	x
Introduction	1
Behavioral Assessment	2
Weaknesses of Behavioral Assessment	7
Direct Observation of Behavior	8
DBO methods	9
DBO psychometrics	16
DBO weaknesses	18
Computerized Behavior Observations	22
CBO formats	24
Psychoeducational Decision Making	27
Sources of Error	30
Accuracy and Direct Behavior Observation	37
Improvements offered by CBO	41
Problems Associated with CBO and DBO	45
Research Questions and Hypotheses	49
Method	52
Participants	52
Materials	53
Testing Room	53
Coded Packets	54
Informed-Consent Forms	55
Participation Forms	55
Partial-Interval Recording Forms	56
Narrative Recording Forms	56
Timer	56
Videotape	57
Video Projection	58
Computers	58
Computer Observation Software	58
Criterion Measure for Accuracy	59
Global Rating Forms	60
Measure of Psychopathology	62
Student Assistants	64
Design	64
Power Analyses	67
Procedure	69
Pre-test Data Collection	69

Introduction to the Study	70
Referral Condition	71
Target Identification Training	72
Rationale for target selection	73
Method Training	74
NM group	75
NAR group	76
PIF group	76
CBO group	77
Pilot Investigation of Training Sequence	78
Observation Accuracy Trial	79
Observation Trial	80
Post-Observation Survey	80
Debriefing	80
Results	81
Analysis 1 - Observational Accuracy	82
Internalizing accuracy measures analysis	84
Externalizing accuracy measures analysis	87
Analysis 2 - Confidence Ratings	93
Internalizing confidence measures analysis	96
Externalizing confidence measures analysis	99
Analysis 3 - Global Ratings	103
Analysis 4 - Psychopathology measures	112
Discussion	117
Effects on Accuracy	117
Effects on Confidence	122
Effects on Measures of Psychopathology	124
Limitations of the Research	126
Limitations related to internal validity	127
Limitations related to construct validity.. ..	128
Limitations related to external validity	130
Implications of the Research	135
Directions for Future Research	137
Summary	139
Appendices	
Appendix A: IRB Approval	141
Appendix B: Web-posting	142
Appendix C: Sign-up sheet	143
Appendix D: Informed consent letter	144
Appendix E: Participation form	145
Appendix F: PIF 2-minute forms	146
Appendix G: PIF 10-minute forms	147
Appendix H: Narrative recording forms	148
Appendix I: Rating form	149
Appendix J: Global rating form	151

Appendix K: CBO software	153
Appendix L: Referral conditions	154
Appendix M: Training sequence for each group	156
Appendix N: Post-observation instructions	177
Appendix O: Debriefing	181

List of Tables

<u>Table 1</u> Univariate source tables for the effect of Method on the internalizing accuracy measures "stupid" and "stomach"	75
<u>Table 2</u> Mean differences between method groups on internalizing accuracy deviation measures	76
<u>Table 3</u> Univariate source tables for the effect of Method and Referral on the externalizing accuracy measures "disagree" and "switch"	79
<u>Table 4</u> Mean differences between method groups on externalizing accuracy deviation measures	81
<u>Table 5</u> Mean differences between referral groups on externalizing accuracy measures	82
<u>Table 6</u> Univariate source tables for the effect of Method and Referral on the internalizing confidence measures "confidence: stupid" and "confidence: stomach"	85
<u>Table 7</u> Mean differences between method groups on internalizing confidence in accuracy ratings	86
<u>Table 8</u> Univariate source tables for the effect of Method and Referral on the externalizing confidence measures "confidence:disagree" and "confidence:switch"	88
<u>Table 9</u> Mean differences between method groups on externalizing confidence in accuracy ratings	90
<u>Table 10</u> Univariate source tables for the effect of method and referral on the global internalizing psychopathology measures	92
<u>Table 11</u> Mean differences between referral groups on global internalizing combination measures	94

<u>Table 12</u> Univariate source tables for the effect of Referral and Method on the internalizing and externalizing psychopathology measures	96

<u>Table 13</u> Mean differences between referral groups on internalizing psychopathology measures	97

List of Figures

<u>Figure 1</u> Mean deviation scores from the criterion scores for each target	83
<u>Figure 2</u> Means of confidence ratings for each Target	94
<u>Figure 3</u> Means of global ratings on the internalizing scales	104
<u>Figure 4</u> Means of global ratings on the externalizing scales	110
<u>Figure 5</u> Means of psychopathology measures (CBCL-DOF) for both internalizing and externalizing items	113

Introduction

Behavioral assessment maintains a focus on an individual's behavior within an environment, is conducive to designing and to monitoring interventions, and is widely used in applied settings. Direct behavior observation is a method of behavioral assessment that requires a minimal degree of inference, and is among the most widely used methods of data collection in psychology (Mash & Terdal, 1988). Despite their popularity and their putative less susceptibility to bias than traditional norm-referenced measures, the complex real-time encoding and tabulating of behavioral data are cumbersome. Recent efforts to enhance the efficiency of direct behavior observation methods have resulted in computer behavior observation programs that allow an observer to collect and to tabulate behavioral data via a mouse click or button press (Kahng & Iwata, 2000). It often is assumed that the improved efficiency offered by a computer behavior observation method yields improved accuracy of an observation, and hence, the integrity of applied decisions and interventions. This assumption of improved accuracy using a computer behavior observation method, however, has not been

researched directly. The present research investigated the relationship between different methods of direct behavior observation assessment, the accuracy of the observations, and the accuracy of the decisions that are made resulting from the different direct behavior observation methods. The literature review that follows explores direct behavior observation as a facet of behavioral assessment, the development and application of computer behavior observation methods, and the relevance of the clinical decision-making literature in using direct behavior observation methods.

Behavioral Assessment

Behavioral assessment is a broad term that encompasses characteristically occurring concepts, methods, and purposes of evaluation that are best understood within the context of behavioral intervention (Mash & Terdal, 1988). Behavioral intervention involves a range of deliberate problem-solving strategies that incorporate a flexible and continuing process of hypothesis testing. Behavioral assessment encompasses a multitude of methodologies including the use of structured and non-structured interviews, behavioral checklists, self-monitoring procedures, analogue

methods, psychophysiological recordings, and direct observations of behavior (Cone, 1978; Mash & Terdal, 1988; Shapiro & Kratochwill, 2000b). In addition to methods of behavioral assessment, Cone (1978) identified modalities (e.g., motor, cognitive, psychophysiological) and generalizations (e.g., scores, items, times, settings, methods, dimensions) as domains incorporated within a behavioral assessment. Given these parameters, behavioral assessment approaches are flexible, and may be applied to many different testing situations.

Behavioral assessment techniques offer a viable alternative to traditional or norm-referenced psychological assessments because data are collected in natural environments with minimal observer inference (e.g., from memory and impressions; Gambrill, 1990; Hintze & Shapiro, 1995). The theoretical assumptions and conceptual underpinnings of behavioral assessment fundamentally differ from the assumptions and theory of traditional assessment procedures. For example, from a traditional assessment perspective, behaviors are seen as symptoms of underlying conditions or traits of an individual. Thus, a goal of traditional assessment is to identify these underlying traits or dimensions from

behavior samples (Mischel, 1968; Skinner, Dittmer, & Howell, 2000a). On the other hand, the goal of behavioral assessment is to determine what environmental influences affect an individual's behavior. Behavioral assessment views behavior as a function of a person's interaction with the environment, and therefore aspires to gather verifiable information about the occurrence of discrete units of behavior.

The methods of behavioral assessment are consistent with behavioral theories that map environmental controls on a behavior. The most basic of these theories is the three-term contingency or ABC model, which poses that behavior (B) is mediated by environmental antecedents (A) and consequences (C) (Bijou, Peterson, & Ault, 1968). Another well-known model, the SORKC model, integrates establishing operations of the organism and environmental contingencies as additional influences on behavior (Kanfer & Phillips, 1970). The acronym "S-O-R-K-C" serves as a way of organizing classes of assessment information into categories of antecedent and consequent events. Each component of this model may be designated as a target for treatment. "S" refers to stimulus, or the external environments that have some functional

relationship to behavior. "O" refers to the biological status of the organism, and includes genetic, physiological, and mechanical variables that influence behavior. "R" refers to the measured response of an organism. "K" refers to the contingency relationships between the response and its consequences (e.g., schedules of reinforcement). Lastly, "C" refers to the consequences of the response (Kanfer & Phillips, 1970; Mash & Terdal, 1988)

Finally, the problem-solving model proposed by Bergan and Kratochwill (1990) is largely representative of behavioral applications to behavior change and to case consultation. This model is divided into four phases of problem solving that lend themselves to behavioral assessment approaches: (a) problem identification, (b) problem analysis, (c) plan implementation, and (d) treatment evaluation (Bergan, 1995; Bergan & Kratochwill, 1990). Any single domain within these models can be manipulated or observed as a target in a behavioral assessment.

Behavioral assessment techniques also differ from traditional assessment measures in terms of the purpose of assessment. Whereas traditional assessment yields

information that is useful for making decisions regarding diagnostic classification and educational placement, behavioral assessment approaches are designed specifically for the purpose of intervention design and maintenance. The primary advantage of norm-referenced assessment is that criteria (e.g., school placement) can be set and measured in standardized ways that are considered to be scientific and fair. In many instances, however, a client's needs are not met with norm-referenced assessment because information about how an individual functions in relation to the natural environment is not made available. Specifically, norm-referenced assessment often does not provide information needed to identify specific behaviors to target for intervention, to develop treatment procedures, nor to monitor effectiveness (Shapiro & Kratochwill, 2000a, 2000b; Rechly & Ysseldyke, 1995). When relying solely on norm-referenced measures, the behavior on which the referral is based may never even be directly assessed (Shapiro, 1987). Behavioral assessment techniques specifically are designed for these purposes (Hintze & Shapiro, 1995).

Finally, traditional standards of test reliability and validity apply differently to behavioral assessment techniques. For traditional measures, variability in data is expected, but is to be kept within certain limits in order for a measure to have integrity. The familiar standards of concurrent validity, construct validity, and internal consistency are all indicators of how well a test can measure constructs that are intrinsic to individuals. For behavioral assessment, however, variability in data is embraced, as it is not regarded as a primary indicator of test integrity. Instead, variability is often a clue that leads investigators or practitioners to identify the factors responsible for individual differences (Cone, 1981).

Weaknesses of Behavioral Assessment

The formulation of a behavioral assessment system precedes the application of the system to monitoring patterns of behavior. Constructing a valid behavioral assessment system that incorporates many methods is not automatic and free of error. This process first involves determining the purpose of the observation. Next, the targets (e.g., what is observed and manipulated) must be identified and clearly defined. Finally, the amount of

behavior to be observed, the minimum information needed to establish the quality of the data to be produced, and available resources for recording the data need to be considered (Foster & Cone, 1995). All of these steps are subject to biasing influences. Specifically, there is literature documenting the tendency of practitioners inadvertently to adopt hypothesis-confirming heuristics in reaction to personal ideologies, referrals, and administrative pressures (Arkes, 1981; Harris, 1994). The construction of a behavioral assessment is subject to these cognitive errors; however, there has not been research suggesting the specific influences of these biases (Shapiro & Kratochwill, 2000b).

Direct Observation of Behavior

The direct observation of behavior is one technique of behavioral assessment that is widely used, so much so that this technique is often mistakenly regarded as being synonymous with behavioral assessment (Mash & Terdal, 1988). Direct behavior observation ideally involves recording behavior when it occurs, using trained and impartial observers, establishing clearly defined targets, structuring methods for recording behaviors, and establishing reliability procedures

(Hintze & Shapiro, 1995). Direct behavior observation is a highly regarded technique because it involves a minimum of inference from observation to data recording. In addition to its low level of inference, direct behavior observation also presents a methodology where there is a low risk for scoring error, observers are easily trained, interrater reliability measures are easily computed, and data can be graphed for interpretation (Hintze & Shapiro, 1995). Moreover, the direct link of observations to data designates direct behavior observation as a highly practical and desirable method to use in applied settings (Shapiro, 1987).

Direct behavior observation is a theoretically relevant method for many other scientific fields including speech/language pathology, physical/occupational science, sociology, and anthropology. In psychology, direct behavior observation can be an effective tool for evaluating the effects of social, educational, and pharmacological interventions (Thompson, Felce, & Symons, 2000). Direct behavior observation has been used effectively in different disciplines of psychology including ethology, experimental psychology, school psychology, industrial

psychology, developmental psychology, social psychology, and applied behavior analysis (Foster & Cone, 1995). In addition to being a behavioral assessment method that can be used to perform functional analyses and can provide information on intervention effectiveness, direct behavior observation also may be used to collect data for inter- and intra-individual base rates (Mash & Terdal, 1988).

Direct behavior observation methods. Collecting direct observation data requires the observation and recording of behaviors. Behaviors have both physical and temporal characteristics. These characteristics include the topography or shape of a behavior; the intensity, frequency, duration, or rate of a behavior; and other temporal characteristics such as the latency between the beginning of a stimulus and the initiation of a behavior (Skinner, Rhymer, & McDaniel, 2000). With the exception of measuring the intensity of a behavior, the other dimensions of behavior can be recorded in narrative and time-structured formats.

A narrative recording of behavior is a written account of observed behaviors, and is useful for forming hypotheses regarding the interdependency among

antecedents, behaviors, and their consequences.

Therefore, narrative recordings often precede more structured time-oriented formats. A narrative recording does not usually specify targets prior to an observation, and observations are not usually plotted by time. Consequently, data obtained by a narrative recording are subject to more inference, and are less specific than information obtained from more structured observational methods. For example, a student may be reported as being "not on task for most of the time" rather than being "not on task 60% of the time." Additionally, written descriptions of behaviors and events are likely to differ across observers. This subsequently makes the establishment of trends, variability, and levels of observed behavior inaccurate (Skinner et. al, 2000). Narrative recording is the only form of direct behavior observation that results in written data rather than numerical data. Because of this uniqueness, and the wide use of narrative recording to establish hypotheses regarding behavior; narrative recording was chosen as a variable condition in this study.

In contrast to narrative recording strategies are methods of direct behavior observation that result in quantitative data. Quantitative data allow one to make comparisons within and across recording sessions and facilitate the accuracy that is essential for data interpretation (Skinner, Dittmer, & Howell, 2000b).

There are many procedures that can be used to record observations directly that result in quantitative data. Quantitative data collection requires observers to define behaviors clearly and then to record observed instances of those behaviors, usually by using a tally-mark and grid system. The various direct behavior observation procedures that result in numerical data include event recording, duration recording, latency recording, and interval recording strategies. A direct behavior observation also may be designed to incorporate the recording of antecedents, consequences, setting variables, and variables intrinsic to the client that may be related to the target behavior (Hintze & Shapiro, 1995).

When event recording is used, each occurrence of a behavior is tallied. Event recording is most successful when observing behavior that has a discrete

beginning and ending. Event recording also yields the most thorough record of the occurrences of a behavior (Hintze & Shapiro, 1995). Because the event recording strategy requires continuous observation, it is not practical for monitoring many behaviors, or for monitoring behaviors that occur at a high rate (Skinner et. al, 2000).

The duration of a behavior also can be recorded. The observer notes the length of time from the beginning of the response until its end. For duration measures to be useful, a behavior must have a distinct beginning and ending, and the starting and finishing points of the behavior must be defined precisely (Hintze & Shapiro, 1995). Duration measures are useful for gauging the severity of behaviors (e.g., time of a tantrum, time out of seat); however, there are some behaviors where duration measures are not possible or practical (e.g., cursing, twitching) because they are discrete.

Latency recording is the measurement of elapsed time between the onset of a stimulus and the initiation of a specified behavior (Cooper, 1987). This form of assessment is useful in determining ABC relationships in the context of functional analysis and reinforcer

assessment. This method is impractical for establishing links between the environment and high frequencies of behavior.

A suitable method for obtaining data on the frequency or rate of behavior simply involves recording each occurrence within a specified time frame; this is referred to as interval recording. Interval recording can be constructed as whole-interval, momentary-time-sampling, or partial-interval formats. Usually, a 5- to 30-minute time period is divided into smaller intervals where behaviors are recorded as occurring at a specific time.

If a behavior is recorded as present only when it occurs throughout an entire time interval, the whole-interval method of recording is being used. This method of recording is appropriate for estimates of duration. Whole-interval time sampling tends to underestimate the time an individual spends engaged in a specific behavior (Skinner, Dittmer, & Howell, 2000b).

Momentary time sampling is a method that requires an observer to note whether a behavior is present or absent only during the moment when a time interval begins. In this technique, the observer would not

record any data during the period between intervals, only what was occurring during the instant of observation. Momentary time sampling is another method that is more appropriate for relatively enduring behaviors. Momentary time sampling is most practical when observers are not available for the entire observation period due to the setting or to fatigue (Hintze & Shapiro, 1995). For example, a teacher may use this technique allowing for momentary sampling of behavior while a lesson is taught. An important disadvantage of this method is that useful data can be lost as observations are not recorded between intervals.

Finally, the partial-interval-form recording method does not require the behavior to be present for the entire interval. Any single or multiple occurrences during the time interval will constitute the interval as being scored as behavior present. Behaviors are marked as occurring or non-occurring on a form where a grid offers space for indicating the time and nature of the behavior (for an example of this grid, see Appendix G). The partial-interval recording strategy can provide acceptable duration estimates of behavior. It is particularly useful for recording non-continuous

behaviors that sporadically appear during the observation session. Partial-interval methods are sensitive to both duration and frequency estimates, are time coded, and are not as laborious as event-recording methods. It is for these reasons that the partial-interval method is perhaps the most used for recording direct behavior observation data (Hintze & Shapiro, 1995). Because the partial-interval method of collecting observational data is practical and so widely used, it also was included as a variable condition in this dissertation.

Direct behavior observation psychometrics. Because direct behavior observation is a form of behavioral assessment, it is founded in behavioral theory. Therefore, the psychometric standards of direct behavior observation also differ from the psychometric standards of traditional assessment approaches. As they apply to traditional measures, test-retest reliability, construct validity, and concurrent validity are of little use for direct behavior observation (Cone, 1981). Rather, direct behavior observation emphasizes content validity (Stroshall & Linehan, 1986) and observational accuracy

that can be measured by inter-observer agreement (Cone, 1981).

The emphasis on content validity is important in constructing adequate behavioral assessment procedures because the focus of assessment is on the functional relationship between behaviors and environments. Content validity is determined by evaluating the test-item content (e.g., the targets) for their representation of a larger behavior domain (Stroshall & Linehan, 1986). For example, if a child has been referred for being aggressive, a representative sample of targets might be hitting, slapping, and kicking, with each target being defined so that each occurrence can be recorded accurately.

Cone (1981) proposed that because behavioral assessment adopts a natural-science philosophy, measurement accuracy should be the primary method for establishing psychometric integrity of behavioral-assessment strategies. Behavior is comparable to matter in motion, and the units of a direct behavior observation are anchored in natural phenomena. Information that is collected by observation is as valid as it is a reflection of natural phenomena. Therefore,

the aim of systematic observations is for properly trained observers to produce identical protocols (Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000) and this is measured by interrater reliability (Suen, 1988).

Direct behavior observation weaknesses. Although direct behavior observation is a highly regarded technique of behavioral assessment, there are criticisms for relying exclusively on direct behavior observation for information because it can be cumbersome and it is not an error-free approach. Direct behavior observation often has been described as being too costly or too inconvenient for some settings. For example, a needed assessment of a child in a classroom might require observation time that could be used for other professional duties. The fact that direct behavior observation may be restricted to specific settings presents another problem. For example, an observation might be conducted in a school whereas the behavior of interest occurs primarily at home.

Despite the low level of inference of direct behavior observation techniques, it is not immune to measurement error. First, the mere presence of an

observer may affect the behavior of interest; this threat to validity is referred to as "subject reactivity" (Shapiro, 1996). Some recommendations to reduce subject reactivity are: (a) to provide individuals with a vague rationale regarding the presence of the observer, (b) to have the observer avoid staring at target individuals, (c) to have the observer make no direct contact with individuals, (d) to use equipment such as one-way mirrors and video cameras, (e) to have the observer sit in an inconspicuous area of the room, (f) to have the observer enter the room before the individual enters, and (g) to gain informed-consent from the parents for permission to conduct an observation that is designed to reduce subject reactivity (Shapiro, 1996). Of course, although these suggestions may reduce subject reactivity, they are unlikely to eliminate it entirely.

Second, an observer might not have the physical or cognitive capacity to record target behavior reliably for a given period of time. Just as the behavior of an individual is not assumed to be stable, the recording behavior of observers also may fluctuate affecting the accuracy of an observation. This fluctuation is

referred to as "observer drift" (Skinner et. al, 2000). In addition to physical and cognitive capacity to record, observer drift can be stimulated by poorly trained observers, poorly defined targets, and poorly constructed data-recording systems. Observer drift usually is identified by having another trained observer collect data simultaneously with the primary observer. Data are then compared across observers to yield a measure of interobserver agreement. Cohen's kappa statistic is a preferred measure of interobserver agreement because it corrects for chance observer agreement (Skinner et. al, 2000; Suen, 1988).

Third, just as the selection of targets is subject to bias in behavioral assessment, so is the selection of targets for a direct behavior observation. The most important consideration preceding a direct behavior observation is the selection and formulation of the target behaviors that are truly useful (Bergan & Kratochwill, 1990). The targets of a behavioral observation can be selected for many reasons. A behavior can become a matter of interest for diagnostic purposes if it is representative of a more basic underlying trait or construct (Goodenough, 1949). In

addition, targets can be selected because manipulating behavior change of the targets would be of practical and social benefit to the person and to others around the individual (Hawkins, 1991). Other criteria to consider for selecting target behaviors are: (a) alter behaviors that are most irritating to the mediator involved, (b) alter behaviors that may be relatively easy to change, (c) alter behaviors that will produce beneficial response generalization, or (d) when responses exist as part of a longer chain, alter behaviors at the beginning of the chain (Hintze & Shapiro, 1995; Nelson & Hayes, 1979).

Despite these guidelines, it is possible for the selection of targets to be biased due to premature and largely subjective hypothesis formation (Arkes & Harkness, 1980; Mash & Terdal, 1988). Therefore, Mash and Terdal (1988) suggest an actuarial approach where targets are selected on the basis of their validity and specificity in relation to a diagnostic category, and of their incremental value to assessing the effectiveness of an intervention.

Fourth, the structure of an observation (e.g., format, number of target behaviors, duration of

observation) might exceed the cognitive and physical capacities of an observer to record behavior accurately. For example, recording frequency data for seven targets is more cognitively taxing than recording partial interval data for three targets. The selection of the structure of the observation must be balanced to fit both the purpose of the evaluation and the cognitive abilities of the observer. The validity of direct behavior observation-assessment formulation involving the selection of targets and the structure of data collection methods remains to be investigated (Kratochwill & Shapiro, 2000b).

Computerized Behavior Observations

A recent development designed to improve the efficiency of direct behavior observation procedures has been the introduction of computer systems to record and to analyze behavioral data (see Kahng & Iwata, 2000 for a review). This development has the potential to reduce error sources of conventional direct behavior observation methods (e.g., methods that require a paper, pencil, and a stopwatch) so that observational accuracy can be enhanced. First, computer behavior observation systems have the potential to facilitate the task of

data collection. Computer systems feature the real-time recording of behavior via a mouse-click or a key press, and bypass conventional partial-interval methods that are cumbersome (Kahng & Iwata, 1998). Computer behavior observation systems also eliminate the need to attend to time intervals and to the placement of tally marks. Thus, the amount of time observing the actual behavior is increased when using a computer method, thereby potentially increasing the accuracy of the observation.

Second, the data from a computer behavior observation system are available in a spreadsheet that can be programmed to tabulate and graph data, thus avoiding another error-prone process that usually is done by a calculator or by manually entering the data on a spreadsheet. Computer-observation systems can be programmed to perform a variety of data-analysis tasks such as calculating interrater reliability estimates, percents of behavior occurrence over time, and time-series modeling and analysis of treatment effects. These features reduce the time-consuming efforts of calculation by hand. In addition, these features reduce the possibility of calculation error, and possibly enhance the integrity of an assessment (Kahng & Iwata,

2000). Finally, the computer behavior observation also allows large amounts of data to be stored in much smaller spaces than in conventional methods, possibly reducing the misplacement of files and enhancing the organization of record storage.

Third, the cognitive-load reductions afforded by a computer behavior observation system (e.g., mouse-click data entry, data coding in real time) enhance an observer's capability to record behavioral contingencies that are far more complex than conventional direct behavior observation methods (Sandman, Touchette, Marion & Bruinsma, 2000). For example, computer behavior observation systems have been used effectively to code transitional data between behaviors (Guess, Roberts, Siegel-Causey, Ault, Guy & Thompson, 1993), the interaction of behavior with different environments (Bakeman & Gottman, 1986), and the effects of medication on behaviors (Sandman et. al, 1993). A benefit from the use of computer behavior observation methods in these studies is that the data can be analyzed and graphed with the aid of a computer (Hall & Oliver, 1997).

Computer behavior observation formats. The programs used to collect direct behavior observation

data have either used a bar-code interface, or "point-and-click" and key-press interface to record data. Bar-code technology interfaces were the first to evolve. A typical system used a laptop or pocket computer, a portable bar-code scanner, and a clip-board with clearly indicated bar-codes representing target behaviors. An observer sweeps the scanner over each target bar-code as corresponding behavior is observed. The data are transferred to a spread-sheet program in the computer so that information regarding the time of occurrence, frequency, and duration of the behaviors can be reviewed. Forney, Leete, and Lindburg (1991) reported bar-code methods of observation to be an improvement over other methods such as paper-and-pencil recording and keyboard entry. In addition, the amount of training time was reduced with this method because observers quickly could refer to descriptions of behavior that were printed beside the bar-code labels.

Despite the valuable advantages discussed by Forney et al. (1991), bar-code observational systems present several disadvantages for observational use. For instance, repeated scans are sometimes necessary to record the behavior, making rapidly occurring behaviors

difficult to record (Eiler, Nelson, Jensen, & Johnson, 1989). Advances in scanning technology have not made high-speed scanning a possibility (Saunders, Saunders, & Saunders, 1994). Bar-code scanning equipment is expensive; in 2000, for example, the cost of a two-scanner installation was approximately \$800 (Sandman et. al, 2000). Finally, the visual presence and auditory signals might influence reactivity more strongly than other methods.

It is conceivable that the difficulties encountered with the bar-code system can be corrected with a modified computer program. For example, a computer can be programmed to display buttons in a window representing each target behavior; an option to display the operational definition of each behavior by each button also can be made available. In addition, a function can be specified to assign keystrokes to the buttons instead of having to click them with a mouse. This function might further allow an observer to focus attention on the individual rather than scanning a computer screen for the appropriate button to click. Observational programs also can be used with palm-top computers that, because of their easy handling and

unobtrusive size, can reduce reactivity (Emerson, Reeves, & Felce, 2000; Sandman et. al, 2000).

Observational methods are ubiquitous in applied settings because of their utility in assessing behavior changes in one or a few individuals. The broad speculation of computer behavior observation methods to improve the capabilities of an observer (Kahng & Iwata, 2000), and subsequently the accuracy of an observation, leads to the question of whether computer behavior observation methods can improve upon the accuracy of the decisions that clinicians make. A review of the literature on clinical decision-making follows to examine the potential contributions that computer behavior observation methods might present to the improvement of clinical practice.

Psychoeducational Decision-making

A body of research suggests that professional judgments regarding a person's mental health or character are vulnerable to biasing influences (Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Errors in clinical judgment have serious implications for professional practice. Inaccurate decisions can negatively affect the course of treatment

because of unwarranted social stigma and self-fulfilling prophecy (Harris, 1994). In addition, the financial cost of inaccurate decisions can be considerable, such as in psychoeducational placement decisions (O'Reilly, Northcraft, & Sabers, 1989).

A potential advantage of direct behavior observation methods in terms of decision-making accuracy is related to their low level of inference and minimal subjective input. Whereas other methods of collecting behavioral data (e.g., norm-referenced measures, projective measures, unstructured interviews) involve procedures that potentially introduce bias (e.g., memory of events, data lost in coding, inappropriate norm reference), direct behavior observation methods minimize these sources of bias. Computer behavior observation methods offer further improvement because the automatic time coding, tallying, and tabulation of data are more efficient than direct behavior observation procedures. Given these improvements, it follows that data collection and decision-making accuracy would be enhanced by using computer behavior observation methods. The decision-making literature (Dawes et. al, 1989; Grove et. al, 2000) and additional cognitive-

science literature (Bargh, 1994) are relevant to evaluating why computer behavior observation methods of data collection may be superior to conventional direct behavior observation methods in making clinical decisions.

The decisions that practitioners make are often influenced by environmental contexts (e.g., referral information, pressures to diagnose) that facilitate predictable and expedient cognitive strategies that decrease judgment accuracy (Harris, 1994; Kahneman, Slovic, & Tversky, 1982; Lerner & Tetlock, 1999). When making a judgment, an overwhelming tendency for practitioners is quickly to adopt and to adhere to an initial hypothesis, and either to ignore or to distort other pertinent information that is at odds with this hypothesis. This tendency is referred to as confirmatory bias (Nisbett & Ross, 1980; Temerlin & Trousdale, 1969) and is so powerful that judges think they can recall information supporting a hypothesis that, in reality, never was presented (Arkes & Harkness, 1980).

Sources of Error

Confirmatory bias is associated with an array of cognitive sets that can be activated by a combination of environmental and internal influences (Bargh & Chartrand, 1999; Nisbett & Ross, 1980). For example, a practitioner might adopt a hypothesis based on a superficial similarity or resemblance between behaviors and diagnostic categories (e.g., illusory correlation; Chapman & Chapman, 1969), that, in turn, triggers a diagnosis to come to mind (e.g., visibility and availability heuristic; Tversky & Kahneman, 1984). In addition, these hypotheses are often in accord with preconceived ideas and stereotypes (e.g., Bargh & Chartrand, 1999; Barnett, Lentz, & McMann, 2000; Nisbett & Ross, 1980).

Mistakes in judgement are generally unintended, unknown to the person making them, and occur automatically (Bargh & Chartrand, 1999). Moreover, non-optimal judgement usually takes precedence over decision-making practices that lead to greater accuracy such as examining base rates, properly assessing covariation, and using actuarial strategies for data integration (Dawes et. al, 1989; Faust & Nurcombe, 1989;

Grove et. al, 2000). Errors in judgment can occur automatically because of our limited capacity to organize an information-rich environment (Bargh & Chartrand, 1999; Kahneman, 1973; Kahneman, Slovic, & Tversky, 1982; Miller, 1956). Thus, Arkes (1981) referred to the occurrence of confirmatory bias as a series of simplifying strategies in which practitioners strive to reach conclusions in the presence of a multitude of data.

There is a substantial body of literature demonstrating confirmatory bias in well-meaning but unsuspecting judges. In a study by Gauron and Dickinson (1969), for example, clinicians identified working diagnoses very early in their first interviews with a new client, often within 60 seconds. During the interviews, questions were selected that focused on the confirmation of the initial hypothesis whereas open-ended questions designed to elicit inconsistent information were avoided.

The selection of hypothesis-confirming questions specifically was demonstrated in a study by Snyder and Swann (1978). In this study, undergraduate students were given personality profiles describing either an

extroverted or introverted person, and they were asked to choose 12 questions from a longer list that best would allow them to test the hypothesis that a target person fit the profile they received. Analyses revealed a strong preference for a hypothesis-confirming strategy. For example, participants chose extroverted questions (e.g., "what would you do to liven up a dull party?") more often in the extroverted condition than in the introverted condition.

Arkes and Harkness (1980) replicated hypothesis-confirming reasoning by having participants decide where a plumbing blockage occurred under different conditions where hypotheses were given or not given prior to the task. These researchers not only found that participants who were given a hypothesis actively sought hypothesis-supporting information, but that they also selectively recalled hypothesis-supporting information and forgot information that was at odds with the original hypothesis.

Practitioners also tend to rely on information that is vivid and available to them, with behaviors that are both available and visible being remembered more often than other behaviors (Tversky & Kahneman, 1973, 1984).

Behaviors become vivid when they are bizarre or are over-represented in the person's environment (e.g., in the media). Vivid behaviors become available in memory, and easily can be factored into professional decisions. For example, after attending a conference on Attention Deficit Hyperactivity Disorder (ADHD), a school-psychologist might consider ADHD more often when evaluating students and seeks to confirm signs of ADHD. In contrast, other information that is not as visible or available may be ignored and not factored into a decision. For example, a student being evaluated for ADHD also might demonstrate signs that are equally valid for a learning disability or an anxiety disorder, but the signs for these conditions may be ignored or underestimated.

Moreover, signs that are vivid actually may be invalid, or illusory. Still, these signs can be integrated into decision-making processes. For example, research on illusory correlation (Chapman & Chapman, 1967, 1969) demonstrated how peoples' prior expectations of perceived relations between variables can bias inferences. In these studies (Chapman & Chapman, 1967, 1969), naïve participants were taught to associate

personality characteristics with human figure drawings and Rorschach (i.e., inkblot) responses. Most participants learned to see what they expected to see, and also overestimated the frequency of the learned co-occurrences. This kind of research suggests that, in an applied setting, a practitioner might selectively attend to vivid signs to substantiate an incorrect hypothesis. For example, a drawing of a weeping willow is vivid as a "sad" theme, which, in turn might be taken as a valid sign for depression when, in fact, there is no demonstrable relationship between these variables.

To fuel the occurrence of confirmatory-bias errors, the practice of psychology presents many environmental incentives for these errors to occur. Error becomes systematic in the context of a goal where judgments are expected to be made under conditions of uncertainty (Bargh & Chartrand, 1999). Pressures to diagnose and provide referral information can initiate hypothesis-confirming strategies early in an assessment, and can result in the formulation of decisions that are easily justified, albeit inaccurate (Lerner & Tetlock, 1999; Snyder & Thompson, 1988). In addition, the exposure to the large amount of information available in many cases

can trigger reliance on improper stereotypes regarding a client (Bargh & Chartrand, 1999). Finally, professional practice involves conferences, workshops, and other associations that avail themselves to vivid and accessible information. The way this information is combined for practical purposes is rarely a subject of these meetings (Meehl, 1973).

Referral information presents an opportunity for a decision-maker to see information that is available and vivid, and to formulate a hypothesis that initiates confirmatory bias. One notable study investigating the effects of referral information on decision accuracy recruited school psychologists to evaluate simulated reports of children. The profiles the psychologists were given were identical; however, the profiles were preceded with different placement considerations. Some of the psychologists received the profile as a "Learning Disability" referral and some received a "Gifted" referral. It was found that weighting and recall of assessment data in the report and classification of the described child all were biased by reason for referral (O'Reilly et. al, 1989).

There is evidence in the cognitive-science literature that a referral might not be necessary to initiate bias; instead, contextual variables or the visible traits of an individual who fits a subjective stereotype may be responsible (Bargh, 1994; Bargh & Chartrand, 1999). In a professional setting where one is inundated with assessment and therapy cases, stereotypes and contextual priming may be relied on more than practitioners suspect. Supporting literature includes how contextual priming of trait concepts changes the perceiver's interpretation of an identical behavior (Bargh, 1994; Temerlin & Trousdale, 1969), and how stereotypes of groups automatically become activated on the mere perception of the distinguishing features of a group member (Bargh, 1994). There is evidence that observed behavior readily can be distorted by pre-conditioned biases on the part of both observers and those being observed (Johnson & Lobitz, 1974; Snyder, Tanke & Berscheid, 1977).

Finally, it is important to state that the unintended operation of confirmatory bias creates the illusion that a decision is based on many sources of information whereas, in reality, very few (e.g., three

or four) sources may be involved (Arkes, 1980; Faust et. al, 1989). This illusion may be harmful in that a false sense of confidence is adopted, making it less likely for a practitioner to use more accurate decision-making strategies (Arkes, 1981).

This research on confirmatory bias and multiple sources of predictable error is sobering, but encouraging research has suggested how decision-making can be improved. Although judgment accuracy cannot be improved merely by telling a clinician about the sources of confirmatory bias (Arkes, 1981), more practical corrective procedures have been suggested. These suggestions include: (a) actively entertain alternate hypotheses, (b) use measures with solid psychometric ratings, (c) assess covariation between signs and condition, (d) use base rates, (e) decrease reliance on memory and the amount of information integrated, and (f) seek out feedback on decisions that are made. (Arkes & Harkness, 1980; Faust, 1986; Harris, 1994).

Accuracy and Direct Behavior Observation

In light of these suggestions, direct behavior observation methods offer promise in improving decision-making accuracy. Specifically, direct behavior

observation provides a direct measure of behavior that requires little inference from the observer and reduces the cognitive load in data collection. In addition, the target-specified structure of a direct behavior observation greatly reduces the possibility of selecting hypothesis-confirming questions or evaluations. Direct behavior observation techniques also can provide specific rates of behavior occurrence that can be used to provide a database for the construction of theories about childhood disorders, and to perform covariation assessment to guide diagnostic or treatment decisions (Mash & Terdal, 1988). Direct behavior observation techniques also provide a record of complex behavior information that is not subject to memory decay. This record can be used to provide feedback regarding the accuracy of a diagnosis or the utility of a treatment decision. Finally, direct behavior observation data may be analyzed statistically to provide predictions of future behavior and to provide an index of intervention effectiveness.

Direct behavior observation techniques can be used both for formal and for applied diagnostic purposes (Mash & Terdal, 1988). For formal diagnostic purposes,

specific categories are assigned to an individual from a system of disease classification such as the DSM-IV (APA, 1994). For applied diagnostic purposes, direct behavior observation is used to understand the nature of a problem, its possible antecedents, treatment options, and outcomes. For each diagnostic purpose, direct behavior observation can facilitate the generation of empirical data to be integrated into clinical decisions. Specifically, direct behavior observation data can determine the validity of targets empirically, can estimate the presence of a target in relation to base-rate occurrences, and can estimate if the target will add incremental validity to other targets and signs (Faust & Nurcombe, 1989).

Direct behavior observation methods also provide structure in the data-collection process so that errors in premature hypothesis confirmation (e.g., asking leading questions during an interview) may be avoided. Harris (1994) suggested that using structured interviews forces a practitioner to ask questions designed to disconfirm hypotheses. Direct behavior observation can be framed as a structured interview where behaviors are

recorded systematically without the element of an interviewer soliciting hypothesis-confirming responses.

A feature unique of direct behavior observation versus other psychometric approaches is that information is not subject to memory decay. More traditional means of assessment, such as interviews, report forms, or checklists, often involve memory decay as a significant source of error. Measures that rely on memory are vulnerable to selective recall where events that support a hypothesis are recalled more often than events that are not (Arkes, 1989; Harris, 1994). The pre-determination of targets and time-based recording yields information that can challenge views simply held in memory (e.g., "I think he was out of his seat all of the time" vs. "He was out of his seat 40% of the time"). Such information lends itself to the construction of an intervention that can alter the frequency of a behavior, and to the evaluation of the intervention's effectiveness.

The opportunity to provide accurate feedback for clinical practice is made available by direct behavior observation techniques. To improve treatment effectiveness, direct behavior observation offers

compelling evidence to countervail initial intervention decisions that otherwise would be maintained. A practitioner is unlikely to evaluate an intervention's effectiveness and to adopt other plans, particularly if confidence in the intervention is high (Einhorn & Hogarth, 1978). Ineffective treatment and diagnostic decisions that are left unchecked can result in self-fulfilling prophecies for the client and practitioner that actually can reverse treatment progress (Harris, 1994). Information provided by direct behavior observation has strong potential to improve practice because such evidence may supersede initial assumptions or hypotheses.

Improvements offered by Computerized Behavior

Observation

Computer behavior observation methods offer additional advantages over direct behavior observation methods toward improving the accuracy and utility of collected information for use in making clinical decisions. Moreover, the strengths of direct behavior observation (e.g., the pre-determined focus of the observation, the memory recall, and the feedback advantages) are also enhanced. First, computer behavior

observation methods present an interface for data collection that reduces the need to concentrate on the physical aspects of recording, and increases the attention devoted to the individual being assessed (Thompson et. al, 2000). It is well documented that cognitive workload negatively affects an individual's capacity to process information observed (Gilbert & Osborne, 1989). The ability of a computer to keep track of time and to locate data points in their appropriate location with a click of a button reduces the cognitive workload of the observer, thus enhancing the accuracy of the observation record. This cognitive-processing reduction represents a substantial improvement over conventional direct behavior observation methods (Kahng & Iwata, 1998). As Fischhoff (1982, p.427) noted, "Just as a mechanically intact airplane needs good instrument design to become flyable, an honest judgment task may only become tractable when it has been restructured to a form that allows respondents to use their existing cognitive skills to best advantage." Thus, to reach peak efficiency and accuracy in an observation record, the data-collection method must be compatible with the observer's cognitive capacity; this is a distinct

advantage that computer behavior observation possesses over conventional direct behavior observation methods.

Second, computer behavior observation methods store a record of the data in the computer where it can be filed, tabulated, graphed, and analyzed efficiently. In this sense, computer behavior observation provides a means of tabulating data for analysis that is efficient and is largely free from error. Computers are becoming a daily tool for scoring and combining the results of conventional psychometric measures (Garb, 2000), and this also is true for computer behavior observation measures. Because computer behavior observation data are stored in a computer, spreadsheet and graphing applications easily can be written to organize and to tabulate the data. The time-consuming tasks of obtaining percentage of occurrence, estimates of duration, interrater reliability, time-series analysis, and plotting the results on a graph can be executed easily by a computer for large amounts of data.

Third, graphs of behavioral data are perhaps the most common and effective way to convey meaning. With a graph, complex ideas can be communicated clearly, precisely, and efficiently (Tufte, 1983). The use of a

computer as a tool to graph data efficiently is a welcome alternative to practitioners who otherwise might use graph paper and rulers, or who might not use graphs at all. Most behavioral data, in journals and in practice, are presented on a graph depicting decipherable slopes representing consistencies in behavior (Shapiro & Kratochwill, 2000). Given that graphs can clarify data (Tufte, 1983), the computer behavior observation method potentially enhances decision-making accuracy with its ability to produce graphs efficiently.

Finally, the use of single-participant statistical analyses, such as ARIMA (i.e., autoregressive and moving-average) modeling has emerged in reaction to managed care accountability standards (Morgan & Morgan, 2001). ARIMA modeling is an analytical procedure where serial dependency among data can be removed by a statistical transformation, allowing for a statistical analysis of trend differences following an intervention (Glass, Willson, & Gottman, 1975). Computer behavior observation data can be read easily by commercially available statistical software packages offering ARIMA analysis. Thus, computer behavior observation can be an

efficient data-collection tool allowing for subsequent statistical analysis.

Problems Associated with Computerized and Direct Behavior Observation

Computer behavior observation and direct behavior observation methods present many desirable attributes that have the potential to circumvent many of the cognitive sets that interfere with accurate clinical judgment. These methods, however, are subject to other sources of bias that need to be discussed. For example, confirmatory bias can occur in both the data collection and interpretation phases of assessment (O'Reilly et. al, 1989).

The assistance of a computer storing data and producing accurate tabulations reduces error during the interpretation phase. It is still possible, however, that practitioners might disregard or rationalize observational results that are inconsistent with an original hypothesis (Arkes, 1981).

During the data-collection process, an observer might be influenced by environmental pressures (e.g., to diagnose or to judge the effectiveness of an intervention) that provide motivation to become overly

generous or restrictive when tallying each behavior (Bargh, 1994). For example, a practitioner who is confident that an intervention will be effective might become more restrictive in tallying occurrences of problematic behavior following the introduction of the intervention. An opposite effect also might occur; high rates of behavior might prime an observer to identify a behavior if its occurrence is in question (Shapiro & Kratochwill, 2000a, 2000b).

Perhaps the strongest potential problem during the data-collection phase is the process of selecting and defining target behaviors for observation. Decisions that are made regarding what data to collect, how to gather these data, and when to stop are subject to biasing influences. There has not been enough research investigating the validity of procedures to select targets for observation (Shapiro & Kratochwill, 2000b). Direct behavior observation methods might be accurate in obtaining target data; however, the validity of what is being measured is based on the target behaviors that were selected. The procedures of selecting and of defining targets, as well as the direct recording of behavior clearly could be influenced by the influence of

professional contexts (e.g., pre-referral information, pressures to diagnose). This weakness previously has been identified, and there have been recent requests for investigations of potential bias and of improved data collection in direct behavior observation approaches to assessment (Shapiro & Kratochwill, 2000b).

Computer behavior observation methods are often less obtrusive than conventional direct behavior observation methods because data may be collected with electronic devices that are not as detectable as a paper, pencil, and stopwatch method. Thus, computer behavior observation methods have the potential to reduce client reactivity. Computer behavior observation does not completely eliminate this concern, however, because the presence of an observer still can elicit some degree of reactivity.

In addition, it has not yet been researched whether computer behavior observation methods are more susceptible to observer drift than direct behavior observation techniques. That is, the efficiency of a computer interface and analysis may not necessarily remove the potential to attend to unrelated phenomena during an observation.

Finally, the efficiency and improvement of computer behavior observation over direct behavior observation methods have the potential to inflate confidence, and perhaps introduce overconfidence in observational data. This increase in confidence might be due to novelty and perceived scientific integrity of computer observation systems that is emerging in the literature (Kahng & Iwata, 2000). There has been no research to date, however; that suggests that observers using computerized systems would report higher levels of confidence than observers using conventional observation methods. The issue of observer confidence is important to consider because it has been demonstrated that confidence generally exceeds accuracy in a judgment, and accuracy generally is not dependent on confidence (Arkes, 1989; Faust, Hart, Guilmette, & Arkes, 1988). Moreover, decision-makers who are quite certain often make errors (Arkes, 1981; Einhorn & Hogarth, 1978; Faust & Nurcombe, 1989; Fischhoff, 1977).

The possible sources of decision error related to computer behavior observation methods have been presented to demonstrate that this technique is not a panacea. However, the overall potential of computer

behavior observation for improvement over conventional direct behavior observation methods is compelling.

Research Questions and Hypotheses

The proliferation of computer-assisted data-collection systems has significantly enhanced behavioral interventions in applied settings. Several properties of computer behavior observation methods, however, have not been established empirically. First, it commonly is assumed that because computer behavior observation methods are more efficient than traditional direct behavior observation formats (i.e., narrative and partial-interval form) of data collection, computer behavior observation methods are more accurate (Kahng & Iwata, 2000). There has been, however, no research to date investigating whether computer behavior observation methods represent a substantial improvement in accuracy over conventional direct behavior observation methods. This leads to the first question in this dissertation: Are observers who use computer behavior observation methods more accurate than observers using conventional direct behavior observation methods? It was predicted that observers using the computer behavior observation (CBO) method would produce more accurate observational

data than observers using a partial-interval form (PIF) method, which, in turn, would be more accurate than a narrative (NAR) data collection method, followed by a "no-method" condition where no recording method was allowed (NM). That is, $CBO > PIF > NAR > NM$. In addition, it was predicted that the effect of method on accuracy would be mediated by a biasing condition (e.g., referral). This interaction would be such that the accuracy of CBO methods would be relatively unaffected by bias, whereas the accuracy of direct behavior observation methods and NAR methods would be more affected by the presence of biasing conditions.

Second, the relationship of observer confidence and the kind of observation method used has not been investigated. This analysis was important to consider, because confidence generally exceeds accuracy in a judgment, and accuracy generally is not dependent on confidence (Arkes, 1989; Faust, Hart, Guilmette, & Arkes, 1988). Thus, the second research question formulated was: Is observer confidence dependent on the method and on the presence of bias? It was thought that CBO methods, because of their efficiency, would result in observers who are more confident in the accuracy of

their data (Kahn & Iwata, 2000). This speculation, however, was insufficient to warrant a directional hypothesis, and therefore, this question was investigated for exploratory purposes.

Finally, because it was speculated that CBO methods would be more accurate than other direct behavior observation methods, it was predicted that persons who used a CBO method would be less susceptible to decision-making bias than persons who used other direct behavior observation methods (Dawes, Faust, & Meehl, 1989). In addition, it also was reasonable to predict that the decisions of observers who used direct behavior observation methods (PIF and NAR) would be less susceptible to referral bias than those decisions of observers who followed no pre-defined observational structure (e.g., no method given; "NM"). Prior to this study, the application of CBO methods to reducing the influence of bias in clinical decision-making had not been studied. Thus, the third research question formulated was: Are the clinical decisions of observers using CBO methods less influenced by the presence of a biasing referral than those decisions based on data generated from PIF, NAR, and NM methods? It was

predicted that the decisions of observers who used the CBO method would be less influenced by bias than the decisions of observers who used the PIF, NAR, and NM methods.

Method

Participants

Prior to recruitment of participants, this study was reviewed by the University of Rhode Island's institutional review board (IRB). The recruitment of participants began only when the approval indicating a minimal risk to participants was approved by the IRB on 2/21/02. A copy of this approval is on file at the University of Rhode Island IRB, and may be seen in Appendix A.

A total of 243 undergraduate students participated in the study, but three participants were excluded from the analysis because they indicated that they had not worn their glasses during the study. Students were able to view general information about the study, and available participation times on the world-wide web. The web-site was designed for students to participate in research for credit. The information that was posted on this site appears in Appendix B. Students signed-up to

participate on a sheet that was posted in University classroom building. There were 24 dates and time slots accommodating groups of 15 participants. Spaces for the names, phone numbers, and e-mail addresses of the participants were provided (see Appendix C).

Data were collected between 3/5/02 and 4/25/02. The mean age of the participants was 19.19 years ($SD = 3.51$). There were 172 freshman, 41 sophomores, 17 juniors, 4 seniors, and 2 graduate students; participants did not specify a class rank. There were 233 participants who were recruited from an undergraduate introductory psychology class. The other participants were recruited from a sophomore-level class in abnormal psychology. Participation in the study provided partial credit for these classes. A majority of the students were not Psychology majors ($n = 213$); there were 24 Psychology majors and 3 participants did not specify their majors.

Materials

Testing Room

One testing room was used as the setting for all the experimental conditions. The testing room was located in a university computer lab that had 20

computers available for student use, a projector screen, and an overhead projector. The seating arrangements comfortably accommodated 20 people. The seats ranged from approximately 7 to 25 feet from viewing the video screen. Participants were instructed as to where the video would be shown, and were then free to sit anywhere they preferred in the room.

Coded Packets

A coded packet was given to each participant. The packet contained two informed-consent forms, a global behavior rating sheet, an Achenbach Child Behavior Checklist - Direct Observation form (i.e., CBCL-DOF; Achenbach & Edelbrock, 1983), and materials that were relevant to the specific group conditions of the experiment. The components of this packet will be discussed following this description of the packet. A code was written on front of the packet that corresponded to each experimental condition and was formatted as a number followed by a letter, and then another number. The first number corresponded with the method conditions such that "1" was assigned to the NM group, "2" to the NAR group, "3" to the PIF group and "4" to the CBO group. The letter corresponded with the

referral condition each group received such that "A" was assigned to the Internalizing group, "B" to the Externalizing group, and "C" to the General Psychopathology group. The participants were unaware to the meaning of these codes. These codes preserved the confidentiality of each participant and assisted in organizing the data for analysis.

Informed-Consent Forms

All participants received an informed-consent form. This form described the study in such a way that variables and hypotheses were not revealed, that minimal risk was involved to them, and that they reserved the right to end their participation at any time (see Appendix D).

Participation Forms

All participants received a participation form where they filled-out the following demographic information: (a) current date, (b) participation identification code (i.e., on the outside of each packet), (c) year in college, (d) major, (e) if participation counted for course credit, (f) an indication of needing glasses or hearing aids, and (g)

if glasses or hearing aids were being worn at the time of the study (see Appendix E).

Partial-Interval Recording Forms

Participants in the PIF group used four interval-recording forms adapted from those used by McComas, Hoch, and Mace (2000). The first three forms were formatted for a 2-minute observation period (see Appendix F). The third form accommodated a 10-minute observation period (see Appendix G). Each minute of observation was displayed in a 6 by 4 grid with 10-second intervals displayed in the left column and the four targets displayed in the top row.

Narrative Recording Forms

Participants in the NAR group were given one sheet of paper that had "Narrative Recording Form" printed on the top, and a place for their name and the date. The paper was otherwise blank for recording their observations any way they wished (see Appendix H).

Timer

The participants in the PIF condition were instructed to keep track of time intervals by viewing a clock on the computer that presented the time in both analog and digital formats. The clocks on the computers

were synchronized to the second prior to the experiment. The clock program was from the Windows '98 software that was installed on all of the computers.

Videotape

A videotape of a 9-year-old Caucasian female was used as the actor for observation for all groups. The videotape was made with an RCA VHS videorecorder on a tripod. The child and her mother, both of whom knew the experimenter, agreed to be recorded and signed a release to this effect. The video was recorded at the child's home. The scene was the child doing schoolwork with her mother. Prior to the recording, the child rehearsed the targets of misbehavior (e.g., disobeying, saying "I'm stupid," switching activities, and complaining of stomach pain) for 30 minutes. The child was cued to engage in each one of these behaviors for at least 15% of the videotaping period. Forty-five minutes of footage were recorded before editing to ensure an adequate sample of the target behaviors. The original footage was edited digitally using Media 100 version 7.0 software so that each target behavior occurred no less than 13.3%, and no more than 22%, of the time in a 10-minute segment. The tape was also spliced into three,

2-minute sections for training observations, with the target behaviors occurring at least one time and no more than three times in each segment.

Video Projection

The video of the child was displayed to the participants using a JVC LCD projector producing a 52-inch image on a screen. The video sound was produced by speakers that were positioned around the room. The videotape was played on an RCA 4-head video machine.

Computers

Twenty Dell Pentium PCs with 18" VGA monitors were used by the CBO group to enter data, and by the PIF group to keep track of time intervals. For the CBO group, each computer was equipped with a 3.5 floppy disk in the drive with the CBO program on it. The computers used Windows'98 software as an operating system.

Computer Observation Software

Participants in the CBO condition used a computer observation program written by the author in Visual Basic 6.0. The interactive window of this program is presented in Appendix K. This program allowed each participant to enter time-coded data by clicking the mouse on buttons that were labeled with the target

behaviors. The interactive window provided a space for the participants to record their participation code. This code allowed the experimenter to categorize data under corresponding experimental conditions while the identities of the participants remained anonymous. The interactive window also presented a start button, a stop button, a duration timer, four mouse buttons labeled with the targets, and a number indicator under each target button. When the "start" button was pressed, the timers indicated the interval and the duration times. The duration of each observation was fixed to reduce participant error in recording. The observations were time-coded and saved automatically by the program in a spreadsheet format. When the observation was complete, the percentage of time engaged in each target behavior was displayed below each target button. The experimenter had the participants start a new observation by first pressing an "ok" button and then pressing the "start" button.

Criterion Measure for Accuracy

To establish a criterion of accuracy for the observation, five graduate students in psychology were trained using the partial-interval form method of

behavior observation recording. The graduate students viewed the videotape and coded their observations on 2/2/02. The majority agreement of these observers was used to establish the criterion by which observational accuracy would be measured. This method of establishing the criterion was chosen because it would balance out the possible inaccuracies of any one observer. In addition, the examiner wanted to test if the target definitions were established well enough to promote acceptable interrater reliability.

Agreement among the five graduate students was computed with individual Kappa coefficients on each target (Cohen, 1960). The overall agreement was acceptable to establish a criterion, $K_{\text{overall}} = .92$. The agreement for the observers measuring each target was also within an acceptable range, with $k_s = .89$ for verbal disagreement, .91 for self labeling as stupid, .91 for task switching, and .95 for stomach complaints.

Global Rating Forms

Before the participants rated the videotape actor with a measure of psychopathology scale (i.e., CBCL-DOF, see below), they completed a global-rating form (see

Appendix J). This procedure was conducted to determine possible influences of the target definitions upon the results of the CBCL-DOF. The global-rating form required participants to rate the videotape actor on internalizing and externalizing behaviors on a 7-point scale. The first two items on the global-rating form asked general questions about internalizing and externalizing behaviors respectively. These items were presented as follows:

Internalizing item: "One way to classify child behavior problems very generally is as 'internalizing.' This is a group of behavior problems showing symptoms of anxiety, depression, and withdrawal. How much did the child in the video demonstrate internalizing behaviors?"

Externalizing item: "Another way to classify child behavior problems very generally is as 'externalizing.' This is a group of behavior problems showing symptoms of hyperactivity, disobedience, and aggression. How much did the child in the video demonstrate externalizing behaviors?"

The order of these two general items was alternated to counterbalance any primacy effects for the content of the first item. The remainder of the questionnaire consisted of six additional items that solicited ratings of specific behaviors (e.g., "How much anxiety did the child in the video demonstrate?"). These six items

alternated between internalizing and externalizing content to minimize a possible order effect bias.

The 7-point scale for rating each item was constructed in a way that "1" represented no presentation of the behavior whereas "7" represented an extreme presence of the behavior. The language of the rating system was constructed to present a range of choices that increased in intensity in relatively equal increments. For example, on the item asking to rate how anxious the child was, the ratings were presented as follows: "1 = No anxiety, 2 = Minimal anxiety, 3 = Some anxiety, 4 = Moderate anxiety, 5 = High anxiety, 6 = Very High anxiety, 7 = Extreme Anxiety."

Measure of Psychopathology

Each participant rated the behaviors of the actor using the Child Behavior Checklist - Direct Observation Form (CBCL-DOF). This measure generated values for internalizing and externalizing psychopathology. This measure was designed for clinicians to rate the occurrence of behaviors on a 4-point scale (e.g., not observed, slight occurrence, definite occurrence with mild intensity and less than 3-minutes duration, definite occurrence with severe intensity and greater

than 3-minutes duration). There were 96 items on this form that have a high degree of overlap with the behaviors rated on the parent form of the CBCL system (Achenbach, 1986). The parent scales have acceptable reliability with internal consistency ratings ranging from .76 to .92. The CBCL-Parent Report Form demonstrates the ability to discriminate reliably between clinical and non-clinical groups (Achenbach, 1991b). The correlations between the CBCL-DOF and the CBCL-Parent Report Form have not been investigated for validity; however, again, there is considerable overlap between the items on both of these forms (Achenbach & Edelbrock, 1983). The CBCL-DOF reports respectable levels of interrater reliability, ranging from .76 to .96 (Achenbach & Edelbrock, 1983).

To score the CBCL-DOF, the internalizing and externalizing items were first isolated based on their factor designation and item-similarity on the CBCL-parent report form (Achenbach & Edelbrock, 1983). T-scores for the internalizing and externalizing composites were not used for this study as a ceiling effect would have occurred (i.e., a item score of "6" or higher on each domain would have produced a T-score

above the 99 range). The raw-score for the internalizing, externalizing, and the combined total of these scores were used as dependent measures in this study.

Student Assistants

Two undergraduate psychology student assistants volunteered to help the experimenter with the study. These students were involved in making the packets for the experiment and scoring the returned packets.

Design

There were four research designs used to provide information regarding the three research questions in this dissertation (i.e., there were two designs that answered one particular question). In all designs, there were two independent variables: (a) the method, and (b) the referral (i.e., biasing condition). In all of the designs the method comprised four levels: (a) a computer behavior observation condition (CBO), (b) a partial-interval form condition (PIF), (c) a narrative condition (NAR), and (d) a no method condition (NM). In addition, all of the designs comprised three levels for the referral: (a) an internalizing referral condition, (b) an externalizing referral condition, and (c) a general

psychopathology condition. The designs were also completely crossed, resulting in 4 by 3 designs comprising 12 possible experimental conditions. The number of participants for each experimental condition ranged from 14 to 26, and averaged 20 participants per condition. A table of the number of participants in each experimental condition appears in Table 1.

Table 1
Participant membership for each experimental condition

	CBO	PIF	NAR	NM
INT	17	22	14	21
EXT	23	18	26	16
GP	18	21	23	21

Total N = 240

In the first design, there were four dependent variables for accuracy on each of the four target behaviors. The dependent measure for accuracy was calculated by taking the absolute value of the difference between the participant's estimate and the

criterion for each of the four targets. This difference score was adopted because it was felt to be the best measure of accuracy comparing the participants' performances with the criterion group.

There were four dependent variables in the second design. These were the confidence ratings on each of the four target behaviors. The rating of confidence was determined by the participants' subjective rating of confidence on a 7-point scale. The points on the scale will be indicated as follows: 1 = Absolutely Uncertain, 2 = Minimal confidence, 3 = Some confidence, 4 = Moderate confidence, 5 = Highly confident, 6 = Very Highly confident, 7 = Absolutely certain.

The third and fourth designs were constructed to answer the research question pertaining to psychopathology measurement. In the third design, there were two dependent variables, the internalizing item ratings and externalizing item ratings on the global rating form.

In the fourth design, there were also two dependent variables that corresponded to measurements of internalizing and externalizing psychopathology. The ratings in the fourth design; however, came from the

internalizing and externalizing raw-scores from the CBCL-DOF.

Power Analyses

A power analysis was conducted on all of the designs factoring in effect sizes within the medium range (e.g., R^2 multivariate = .06) effect size, and alpha level of $p < .05$, and 20 participants per group.

For the 4 x 3 MANOVA with two dependent variables, power was computed for two main effects. The first main effect, the one with 4 levels, had strong predicted power ($P = .841$; R^2 multivariate = .06; $N = 240$). The second main effect, the one with 3 levels, also indicated acceptable power ($P = .89$; R^2 multivariate = .06; $N = 240$). Finally, the power for the interaction was also computed to be within the acceptable limits ($P = .76$; R^2 multivariate = .06; $N = 240$). These power levels were considered acceptable to conduct a study in behavioral research (Keppel, 1991).

Power also was calculated for the follow-up ANOVAs in the two dependent variable designs assuming an alpha-level of $p < .05$, a medium effect size (i.e., R^2 multivariate = .06), and 20 participants per group. Power for the main effect with 4 levels was predicted to

be .91, power for the main effect with 3 levels was predicted to be .95, and power for the interaction was predicted to be .83. These power levels were also judged to be acceptable for the behavioral sciences (Keppel, 1991).

Finally, a power analysis was also conducted for the designs that incorporated four dependent variables. However, this analysis was originally conducted within the parameters of a 3 by 3 design because an earlier proposal for this study had planned for such a design. The two main effects had acceptable predicted power ($P = .74$; R^2 multivariate = .06; $N = 180$). The power for the interaction was also acceptable ($P = .61$; R^2 multivariate = .06; $N = 180$). For the follow-up ANOVAs, power for both main effects was predicted to be .85, and power for the interaction was predicted to be .75. These estimates would have been lower had they been calculated for a 4 x 3 design. However, given these moderate power ratings, the experimenter decided that 240 participants with 20 participants per group would be an acceptable sample size for the study.

Procedure

Prior to the study, a criterion observation was established by having five graduate psychology students perform an observation on all of the video segments. A kappa coefficient was selected here because it has been found to be a more robust measure of agreement than percentage-agreement indices (Cohen, 1960; Suen & Lee, 1985).

The data points were collapsed into a criterion by keeping data points where there was a 60% agreement or better. This established criterion was used to train the participants in the first two training sessions. This criterion was also used to judge the accuracy of each participant's observation.

Pre-test Data Collection

Participants were recruited and randomly assigned to one of the 12 experimental conditions. There were 48 slots of time where the experiment was to run, thus allowing for 4 opportunities to run each experimental condition.

The participants were blind to the hypotheses of the study. Each participant was greeted by the examiner outside the door of the testing room and was given a

packet containing the two informed-consent forms (Appendix D), the participation form (Appendix E), the accuracy and confidence rating form (Appendix I), the global-rating form (Appendix J), and the CBCL-DOF form. Participants in the NAR group received an additional narrative recording form (Appendix H). Participants in the PIF group received three 2-minute (Appendix F), and one 10-minute recording sheets (Appendix G). Participants who indicated that they had a sensory impairment with no corrective devices were not counted in the data. A period of 5 minutes was given for any participant who arrived late to enter the room. After this period, the door was closed and a "do not disturb--testing in progress" sign was placed on the door.

Introduction to the Study

Once the participants entered the room, the informed-consent form was reviewed by the examiner. The participants were then asked to sign both informed-consent forms, one of the forms was returned to the examiner. The participants were instructed to keep the other form for their records. The participants were then asked to fill out the participant form and place it in their packet. The examiner began to follow a script

of instructions. This script first welcomed the participants to the study and then briefly described the video-tape actor in a way that varied with the referral condition (see Appendix M). The participants were then trained in target identification and in method of observation. The script that the examiner followed for this training appears in Appendix M.

Referral Conditions

There were three referral conditions that the participants were assigned to: (a) an internalizing referral, (b) an externalizing referral and, (c) a general psychopathology referral. The referral information was received once by each group, and each referral was constructed as such:

"This child has been referred to the school psychologist because her teacher is concerned about her behavior. [insert: biasing condition] The purpose of your observation is to gain information that will be used in planning an intervention for this child that can be used by the teachers in her school."

The participants in the general pathology referral group received no biasing information in the insert. The participants in the internalizing referral group were told in the insert "Specifically, this girl has demonstrated anxiety and depression, and has complained

about body aches and pains." The participants in the externalizing group were told in the insert "Specifically, this girl has demonstrated difficulty with attention and following directions, and has also demonstrated aggression." The script for the referral conditions is presented in Appendix L.

Target Identification Training

All groups were trained in target identification in two phases. In the first phase, the targets were labeled and defined orally by the examiner. In the second phase, a series of three 2-minute video segments were shown where instruction and feedback were given in identifying the targets.

The four targets were labeled as: (a) verbal disagreement, (b) saying "I'm stupid," (c) switching activities, and (d) complaining of stomach problems. "Verbal Disagreement" was defined as "any time the girl in the video says 'no,' says that she does not want to do the work, or says that she wants to do something other than her work." "I'm Stupid" was defined as "anytime the girl in the video says 'I'm stupid,' 'I'm so stupid,' 'I don't know,' or 'I can't do this.'"

"Switching activities" was defined as any instance where

the girl in the video "gets up out of her seat, plays with items in the basket, plays with the dog, or plays with the soda can." Finally, "Stomach Problems" was defined as when "the girl says 'my stomach hurts' or 'I need medicine.'"

During the second phase of target-identification training, all of the participants viewed three 2-minute video training segments. In the first segment, the target behaviors were identified by the examiner as they occurred. In the second segment, the participants independently conducted the observation and received feedback regarding the accuracy of their observations. Finally, the third segment served as a rehearsal observation and involved no target identification or performance feedback to the participants

Rationale for target selection. The four targets (i.e., verbal disagreement, saying "I'm stupid", switching activities, and complaining of stomach problems) were selected because they corresponded with behaviors that have shown high factor loadings on either the internalizing or externalizing factors of the CBCL, and low loadings on the opposite factor. For example, "verbal disagreement" corresponds with "disobedient" on

the CBCL. "Disobedient" loads .64 on the Externalizing scale for 2-factor Varimax rotation in the girls (age range 6 - 11) sample. The other externalizing target, "switches activities" is analogous to "can't concentrate" which loads .65 on the externalizing scale. The internalizing targets also were selected for their discriminative factor loadings. To call oneself "stupid" was considered as comparable to "feels worthless," which loads .66 on the internalizing scale. Finally, "stomach problems" loads .60 on the internalizing scale. None of these above items were found to load more than .30 on the opposite scale, thus demonstrating that these targets were orthogonal representations of the internalizing and externalizing scales (Achenbach & Edelbrock, 1983).

Method Training

There were four method groups that comprised the second independent variable in this study (i.e., NM, NAR, PIF, and CBO). Each of these four groups received different training on how to record behaviors using their respective method. This training was integrated with the target-identification training. The method-training process progressed in three parts: (a) a

demonstration trial, (b) a feedback trial, and (c) a practice trial. In the demonstration trial, the methods were described, demonstrated, and then practiced when viewing a 2-minute video segment. The target identification and recording in this demonstration was based on the criterion data from the graduate student observers. Each target behavior was clearly indicated by the experimenter when this 2-minute video segment was viewed. In the feedback trial, participants recorded behaviors for a 2-minute video segment and were given feedback on the accuracy of their recordings that were also based on the criterion data. The feedback identified the number of occurrences for each of the target behaviors; this feedback was given following the second training observation. Finally, the practice trial was a 2-minute observation where no instruction or feedback was given.

NM group. Participants in the NM group (i.e., no method given) received no training on recording behavior and were not allowed to record observations using paper and pencil. These participants viewed the three video training segments. The experimenter indicated the targets as they occurred in the first segment, mentioned

that each of the targets occurred following the second segment, and did not provide any feedback following the third segment.

NAR group. Participants in the NAR group were instructed to record their observations using paper and pencil on a "Narrative Recording Form." This form had a space for the participant's personal code and instructions to "record what you see on the video." The paper was otherwise blank (see Appendix H). During the demonstration trial, the experimenter demonstrated a method of recording the observations by writing the names of the targets as they occurred. The experimenter's demonstration was projected on the wall beside the screen by an overhead projector. In the feedback trial, the experimenter indicated that all of the behaviors occurred and that there should be a written record of these behaviors on the target form. A sample form, with all target behaviors written, was shown on the overhead projector.

PIF group. Participants in the PIF group used an interval recording form adapted from those used by McComas, Hoch, and Mace (2000, pp.118-120). This method allows for both the time and the kind of behavior to be

recorded. A behavior is indicated as occurring by placing a check mark under the appropriate target column and in the appropriate time-interval row. Participants in the PIF group were given three training forms that were formatted for a 2-minute observation (see Appendix F). They also were given a 10-minute recording form (see Appendix G). Participants were instructed to keep track of time intervals by viewing the group-synchronized clock that was displayed on the computer in front of them.

The experimenter used an overhead projector to demonstrate the PIF method to the participants. A completed form displaying the appropriate interval-target occurrences was shown for feedback following the second trial.

CBO group. Participants were seated at a computer where they also could view both the computer screen and the video projected on a screen in front of the room. The participants were instructed to log their participation code and the date on indicated boxes on the computer. The examiner read instructions as to how to use the computer for an observation (see Appendix M). These instructions specified that a behavior be recorded

by clicking on one of the four target buttons once. Each mouse-click was to correspond with one occurrence of the behavior. In addition, clicking the "undo" button allowed the participants to erase a previous unintended indication of a behavior. The participants were told that following their observation, the percent of time that the child in the video engaged in each behavior would be displayed under each target button.

In the demonstration trial, the participants recorded behaviors, along with the experimenter, as a primer to understand the CBO method. It was indicated to the participants that the percentage of time engaged in each behavior would be displayed on the screen following each 2-minute trial. This was done so that the participants could verify if they were correctly recording the targets. Following the first two, 2-minute trials the experimenter announced the correct percentage of engaged behaviors to the group.

Pilot Investigation of Training Sequence

A pilot study was conducted with two graduate students in a non-psychology program to see if this four-step training sequence was adequate to understanding the PIF and CBO methods. Both students

demonstrated agreement on all but one data point for the two observations (i.e., 98%) and both students related that they understood the procedure following the training period.

Another pilot study was conducted with five graduate school-psychology students and a faculty member serving as observers for a 10-minute PIF observation. There was a 93% agreement on the data of these observers. There was, however, a suggestion that the PIF form should incorporate the time intervals within the form to enhance understanding. This change was made to the PIF form, which can be seen in Appendix G.

Observation Accuracy Trial

Following the demonstration and feedback trials, participants in all groups were instructed to observe a 2-minute video segment. All participants were told that feedback would not be given on this trial. This procedure served as a practice session for participants using each respective method. The 2-minute footage for this trial contained one occurrence of each target behavior.

Observation Trial

Following the training trials, the participants observed a 10-minute video of the same child. They recorded the behaviors using the method by which they were trained. None of the groups was given feedback for this trial. The data from this trial was used to measure observational accuracy in this study.

Post-Observation Survey

Following the observation trial, the participants were given instructions to respond to the survey forms that were in their packets. A script of these instructions appears in Appendix M. In these instructions, the participants in all groups were asked to fill out the accuracy and confidence rating form (Appendix I), the global rating form (Appendix J), and then the CBCL-DOF form, that were in their packets.

Debriefing

Following the study, the participants were debriefed. The experimenter was careful not to reveal the other conditions or hypotheses because knowledge of these conditions might have been communicated to other students and would have potentially biased the results of the study. The participants were told that they

could obtain information regarding the study hypotheses and results following the data collection period. (see Appendix O).

Results

To investigate each research question, four 4 x 3 between-subjects multivariate analyses of variance (MANOVA) tests were performed using SPSS version 10.1. Each MANOVA investigated the effects of Method (4 levels) and Referral (3 levels) on each of the following clusters of dependent variables: (a) observation accuracy score (i.e., Analysis 1), (b) confidence rating in observation accuracy (i.e., Analysis 2), (c) global behavior rating (i.e., Analysis 3), and (d) CBCL-DOF scores for measuring psychopathology (i.e., Analysis 4). For each analysis, the dependent-variable clusters were partitioned into internalizing and externalizing components to avoid redundancy among the variables on these factors. Significant effects for either Method, Referral, or their interaction on the dependent-variable clusters were followed up with univariate analyses of variance (ANOVA) on each of the individual dependent variables. Finally, Tukey's Honestly Significant Difference (HSD) tests were used to investigate all of

the possible post-hoc mean differences at $\alpha_{FW} = .05$. The Tukey HSD was chosen because of its properties in controlling for familywise error-rate across all possible mean comparisons (Keppel, 1991).

Analysis 1 - Observation Accuracy

To investigate the effects of Method and Referral on observation accuracy, a 4 x 3 between-subjects MANOVA was performed on each set of internalizing (e.g., calling oneself "stupid", and complaining of stomach problems) and externalizing (e.g., verbal disagreement, and switching activities) accuracy measures. The accuracy measures were determined by the absolute value of the difference between the participant's response and the criterion established by the five graduate-student observers. The independent variables were type of method (e.g., no method, NM; narrative recording, NAR; partial-interval form, PIF; computer-based observation, CBO), and type of referral (e.g., general pathology, internalizing information, externalizing information). Graphs of the mean deviation from the criterion scores for each target are presented in Figure 1.

Figure 1

Mean deviation scores from the criterion scores for each target

Figure 1, continued

Internalizing accuracy measures analysis. With the use of Wilks' criterion, the combined internalizing accuracy measures were significantly affected by Method,

$F(6, 454) = 16.89$, $p < .001$; but not by Referral, $F < 1$; and not by their interaction, $F < 1$. These results reflect a strong association between Method and the combined internalizing accuracy measures, partial $\eta^2 = .18$.

The significant main effect for Method on internalizing accuracy was further analyzed with separate univariate ANOVAs on each specific internalizing measure; "self-labeling as stupid" (i.e., "stupid") and "stomach complaints" (i.e., "stomach"). In these analyses, Method was found to affect the measurement accuracy of both "stupid," $F(3, 228) = 29.82$, $p < .001$; and "stomach," $F(3, 228) = 31.07$, $p < .001$. The source tables for both of these analyses are presented in Table 2.

Table 2

Univariate source tables for the effect of Method on the internalizing accuracy measures "stupid" and "stomach"

Dependent Variable = self-labeling as "stupid"

Source	df	SS	MS	F	p
Method	3	18091.59	6030.53	29.83	.0001
Referral	2	544.06	272.02	1.35	.26
Method x Ref	6	720.86	120.14	.59	.74
Error	228	46100.32	202.19		

Dependent Variable = stomach complaints

Source	df	SS	MS	F	p
Method	3	16508.23	5502.75	31.07	.0001
Referral	2	360.08	180.04	1.01	.36
Method x Ref	6	961.18	160.20	.91	.49
Error	228	40375.87	177.09		

Post-hoc comparisons between method groups on each specific internalizing accuracy measure were conducted using the Tukey a. In this analysis, the CBO and PIF methods were consistently more accurate than the NM and NAR methods. There were no significant differences in

accuracy between the CBO and PIF groups or between the NM and NAR groups on either internalizing measure.

These results are presented in Table 3.

Table 3

Mean differences between method groups on internalizing accuracy deviation measures

Method	NM	NAR	PIF	CBO
Dependent Variable = self labeling as "stupid"				
NM	--	2.4	17.9*	19.6*
NAR		--	15.5*	17.2*
PIF			--	1.8
CBO				--
Dependent Variable = stomach complaints				
NM	--	5.2	18.7*	19.7*
NAR		--	13.3*	14.4*
PIF			--	1.1
CBO				--
* $p < .001$				

Externalizing accuracy measures analysis. The combined externalizing accuracy measures were affected by both Method, $F(6, 454) = 24.81$, $p < .001$; and

Referral, $F(4, 454) = 3.61$, $p < .01$; but not by their interaction, $F(12, 454) = 1.43$, $p > .05$. The results reflected a strong association between method and the combined externalizing accuracy measures, partial $\eta^2 = .25$. The association between referral and the externalizing accuracy measures was mild to moderate, partial $\eta^2 = .031$.

Although the MANOVA showed no significant interaction between Method and Referral on the two externalizing measures combined, an exploratory analysis revealed a significant interaction for the "verbal disagreement" accuracy measure, $F(6, 228) = 2.23$, $p < .05$. This interaction demonstrated a moderate association, partial $\eta^2 = .06$. This analysis was conducted to follow-up on the hypothesis proposed of an interaction of Method and Referral on accuracy. The simple effects of this interaction revealed no differences between referral groups within the NM condition, $F(2, 55) = 2.66$, $p > .05$; the PIF condition, $F(2, 58) = 1.79$, $p > .05$; or the CBO condition, $F(2, 55) = 2.12$, $p > .05$. There were significant differences, however, between the referral groups within the NAR

condition, $F(2, 60) = 4.92$, $p < .05$. The significance of this contrast revealed a moderate effect size, partial $\eta^2 = .14$. A post-hoc comparison on this effect revealed that participants who were given an externalizing referral were more accurate than those participants who were given a general psychopathology referral, Tukey a ($M_{Diff} = 19.21$), $\alpha_{FW} < .05$.

The significant main effects for both Method and Referral on externalizing accuracy were further analyzed with separate univariate ANOVAs on each specific externalizing measure; "verbal disagreement" (i.e., "disagree") and "switching activities" (i.e., "switch"). In these analyses, Method was found to affect the measurement accuracy of both "disagree," $F(3, 228) = 36.13$, $p < .001$; and "switch," $F(3, 228) = 37.87$, $p < .001$. Although referral evidenced a main effect on "disagree," $F(2, 228) = 6.91$, $p < .001$, no such effect was observed on "switch," $F(2, 228) = 2.70$, $p > .05$. The source tables for these two analyses are presented in Table 4.

Table 4

Univariate source tables for the effect of Method and Referral on the externalizing accuracy measures "disagree" and "switch"

Dependent Variable = verbal disagreement

Source	df	SS	MS	F	p
Method	3	33400.65	11133.31	36.13	.0001
Referral	2	4257.58	2128.79	6.91	.001
Method x Ref	6	4121.35	686.84	2.23	.041
Error	228	70259.85	308.16		

Dependent Variable = switching activities

Source	df	SS	MS	F	p
Method	3	26765.48	8921.83	37.87	.0001
Referral	2	1272.82	636.41	2.70	.07
Method x Ref	6	1547.12	257.85	1.10	.37
Error	228	53711.32	235.58		

Post-hoc comparisons between method groups on the two externalizing accuracy measures were conducted using the Tukey a. In this analysis, CBO and PIF methods were more accurate than NM and NAR methods when measuring

"switch." The CBO and PIF methods were superior to the NM and NAR methods when measuring "disagree," and the NAR method was more accurate than the NM method when measuring "disagree." There was no difference between the CBO and PIF methods when measuring "disagree." These results are presented in Table 5.

Table 5
Mean differences between method groups on externalizing accuracy deviation measures

Method	NM	NAR	PIF	CBO
Dependent Variable = verbal disagreement				
NM	--	11.0*	27.4**	29.3**
NAR		--	16.5**	19.0**
PIF			--	2.5
CBO				--
Dependent Variable = switching activities				
NM	--	5.2	18.7**	19.7**
NAR		--	13.3**	14.4**
PIF			--	1.1
CBO				--
*	$p < .01$			
**	$p < .001$			

Post-hoc comparisons between referral groups on the two externalizing accuracy measures were conducted using the Tukey a. In this analysis, participants receiving the externalizing referral were more accurate for "verbal disagreement" than participants receiving the internalizing referral and the general psychopathology referral. In addition, participants receiving the externalizing referral were more accurate for "switching behavior" than participants who received the general psychopathology referral. These results are presented in Table 6.

Table 6

Mean differences between referral groups on externalizing accuracy deviation measures

Referral	INT	EXT	GP
Dependent Variable = verbal disagreement			
INT	--	8.82**	2.00
EXT		--	10.81**
GP			--
Dependent Variable = switching activities			
INT	--	3.99	1.79
EXT		--	5.78*
GP			--
* $p < .05$			
** $p < .01$			

Analysis 2 - Confidence Ratings

The effects of Method and Referral on observer confidence were investigated with a 4 x 3 between-subjects MANOVA on each set of internalizing and externalizing confidence scales. With the use of Wilks' criterion, the effect of Method on the combined internalizing confidence ratings was significant, $F(6,$

454) = 6.31, $p < .001$. Neither the effect of referral on the combined internalizing confidence ratings, $F < 1$; nor the interaction between Method and Referral, $F(12, 454) = 1.47$, $p > .05$; however, were significant. The significant effect for Method reflected a moderate association between Method and the combined internalizing confidence ratings, partial $\eta^2 = .07$. Graphs of the means for confidence ratings for each target are presented in Figure 2.

Figure 2

Means of confidence ratings for each target

Figure 2, continued

Figure 2, continued

Internalizing confidence measures analysis. The significant main effect for Method on the internalizing confidence ratings was further analyzed with separate univariate ANOVAs on each specific internalizing confidence measure; confidence in the accuracy rating for calling oneself "stupid" (i.e., "confidence: stupid") and confidence in the accuracy of identifying stomach complaints (i.e., "confidence: stomach"). In these analyses, Method was found to affect both "confidence: stupid," $F(3, 228) = 11.01, p < .001$; and "confidence: stomach," $F(3, 228) = 11.31, p < .001$. The

source tables for both of these analyses are presented in Table 7.

Table 7

Univariate source tables for the effect of Method and Referral on the internalizing confidence measures "confidence: stupid" and "confidence: stomach"

Dependent Variable = confidence in identifying actor
self-labeling as "stupid"

Source	df	SS	MS	F	p
Method	3	33.99	11.33	11.30	.0001
Referral	2	3.28	1.64	1.60	.21
Method x Ref	6	11.05	1.84	1.80	.10
Error	228	234.63	1.03		

Dependent Variable = confidence in identifying stomach
complaints

Source	df	SS	MS	F	p
Method	3	38.00	12.67	11.01	.0001
Referral	2	1.17	.58	.52	.60
Method x Ref	6	11.75	1.97	1.75	.11
Error	228	255.37	1.12		

A post-hoc analysis revealed that participants using the CBO method rated their confidence higher than NM participants and NAR participants for both the "confidence: stupid" and "confidence: stomach" targets, Tukey $\underline{a}$, $\alpha_{FW} < .05$. The CBO method did not reveal higher confidence ratings than the PIF method, Tukey $\underline{a}$, $\alpha_{FW} < .05$. The results of this analysis are presented in Table 8.

Table 8

Mean differences between method groups on internalizing confidence in accuracy ratings

Method	NM	NAR	PIF	CBO
Dependent Variable = "confidence: stupid"				
NM	--	.23	.35	.76***
NAR		--	.58**	.99***
PIF			--	.40
CBO				--
Dependent Variable = "confidence: stomach"				
NM	--	.23	.37	.83***
NAR		--	.60*	1.05***
PIF			--	.46
CBO				--
* $p < .05$ ** $p < .01$ *** $p < .001$				

Externalizing confidence measures analysis. An analysis to investigate the effects of Method and Referral on the externalizing confidence ratings was also conducted. With the use of Wilks' criterion, the effect of Method on the combined externalizing

confidence ratings was significant, $F(6, 454) = 3.78$, $p < .001$. Neither the effect of Referral on the combined externalizing confidence ratings, $F < 1$; nor the interaction between method and referral, $F(12, 454) = 1.07$, $p > .05$; however, were significant. The significant effect for Method reflected a mild to moderate association between Method and the combined externalizing confidence ratings, partial $\eta^2 = .05$.

The significant main effect for Method on the externalizing confidence ratings was further analyzed with separate univariate ANOVAs on each specific externalizing confidence measure; confidence in the accuracy rating for verbal disagreement (i.e., "confidence: disagree") and confidence in the accuracy of identifying switching activities (i.e., "confidence: switch"). In these analyses, Method was found to affect both "confidence: disagree," $F(3, 228) = 3.61$, $p < .05$; and "confidence: switch," $F(3, 228) = 5.99$, $p < .001$. The source tables for both of these analyses are presented in Table 9.

Table 9

Univariate source tables for the effect of Method and Referral on the externalizing confidence measures "confidence: disagree" and "confidence: switch"

Dependent Variable = confidence in accuracy rating for
verbal disagreement

Source	df	SS	MS	F	p
Method	3	11.16	3.72	3.61	.01
Referral	2	1.25	.62	.60	.55
Method x Ref	6	9.52	1.59	1.54	.17
Error	228	235.10	1.03		

Dependent Variable = confidence in identifying switching activities

Source	df	SS	MS	F	p
Method	3	22.70	7.57	5.99	.001
Referral	2	.99	.49	.39	.68
Method x Ref	6	6.74	1.12	.89	.50
Error	228	288.14	1.26		

A post-hoc analysis of the revealed that
participants using the CBO method rated their confidence

higher than PIF participants for "confidence: disagree" targets, Tukey $\underline{a}$; $\underline{M}_{diff} = .54$, $\alpha_{FW} < .05$. Participants in the CBO group rated their confidence higher than PIF, NAR, and NM groups for "confidence: switch" targets, Tukey $\underline{a}$, $\alpha_{FW} < .05$. The results of this analysis are presented in Table 10.

Table 10

Mean differences between method groups on externalizing confidence in accuracy ratings

Method	NM	NAR	PIF	CBO
Dependent Variable = "confidence: disagree"				
NM	--	.25	.34	.21
NAR		--	.09	.45
PIF			--	.54*
CBO				--
Dependent Variable = "confidence: switch"				
NM	--	.05	.01	.69**
NAR		--	.06	.74**
PIF			--	.68**
CBO				--
*	$p < .05$			
**	$p < .01$			

Analysis 3 - Global Ratings

Two 4 x 3 between-subjects MANOVAs were used to investigate the effects of Method and Referral on global internalizing and externalizing measures of psychopathology.

With the use of Wilks' criterion, this analysis revealed that the internalizing global measures were significantly affected by Referral, $F(8, 450) = 2.00$, $p < .05$; but not by Method, $F < 1$; and not by their interaction, $F < 1$. These results reflected a mild to moderate association between Referral and the combined global internalizing measures, partial $\eta^2 = .037$. Graphs of the global rating means for the internalizing scales are presented in Figure 3.

Figure 3

Means of global ratings on the internalizing scales

Figure 3, continued

Figure 3, continued

The effect of referral on global internalizing measures was analyzed more closely with 4 factorial ANOVAs on each internalizing global variable: (a) global internalizing combination, (b) global anxiety, (c) global depression, and (d) global withdrawal. These analyses revealed no significant effect for referral on global anxiety, $F < 1$; global depression, $F(2, 228) = 4.25$, $p > .05$; or global withdrawal, $F(2, 228) = 3.78$, $p > .05$. There was, however, a significant effect found for referral on the global internalizing combination, $F(2, 228) = 7.00$, $p < .05$. The source tables for these analyses are presented in Table 10.

Table 11

Univariate source tables for the effect of Method and Referral on the global internalizing psychopathology measures

Dependent Variable = global internalizing combination

Source	df	SS	MS	F	p
Method	3	3.16	1.05	.59	.62
Referral	2	14.00	7.00	3.93	.02
Method x Ref	6	6.75	1.13	.63	.71
Error	228	406.52	1.78		

Dependent Variable = global anxiety

Source	df	SS	MS	F	p
Method	3	2.99	.997	.60	.61
Referral	2	.43	.22	.13	.88
Method x Ref	6	8.40	1.40	.85	.54
Error	228	377.55	1.66		

Table 11, continued

Dependent Variable = global depression

Source	df	SS	MS	F	<u>p</u>
Method	3	5.57	1.86	1.20	.31
Referral	2	8.50	4.25	2.74	.07
Method x Ref	6	4.34	.72	.47	.83
Error	228	354.12	1.55		

Dependent Variable = global withdrawal

Source	df	SS	MS	F	<u>p</u>
Method	3	2.04	.68	.35	.79
Referral	2	7.57	3.78	1.95	.15
Method x Ref	6	5.93	.99	.51	.80
Error	228	442.24	1.94		

A post-hoc analysis with the Tukey a on the global internalizing effect revealed that participants given an internalizing referral tended to rate the actor higher on internalizing than participants who were given a general psychopathology referral, Tukey a $M_{diff} = .68$, α_{FW}

< .05. The results of this analysis are presented in Table 12.

Table 12

Mean differences between referral groups on global internalizing combination measures

Referral	INT	EXT	GP
INT	--	.33	.61*
EXT		--	.28
GP			--
* $p < .05$			

A Wilks' criterion was also the basis for analyzing the effect of Method and Referral on global externalizing measures. This analysis revealed that the global externalizing measures were neither affected by Referral, $F < 1$; Method, $F(12, 595) = 1.68$, $p > .05$; nor by their interaction, $F(24, 786) = 1.30$, $p > .05$. Graphs of the global externalizing means for each target are presented in Figure 4.

Figure 4

Means of global ratings on the externalizing scales

Figure 4, continued

Analysis 4 - Psychopathology measures

To investigate the effects of Method and Referral on a rating of psychopathology (e.g., CBCL-DOF), a 4 x 3 between-subjects MANOVA was conducted on the combined internalizing and externalizing item scores. Raw scores on the CBCL-DOF were used for this analysis because T-score conversions would have introduced a ceiling effect rendering these results uninterpretable. With the use of Wilks' criterion, the psychopathology measures were significantly affected by Referral, $F(4, 454) = 3.44$, $p < .05$; but not by Method, $F < 1$; and not by their interaction, $F < 1$. The results reflected a mild to moderate association between Referral and the combined psychopathology ratings, partial $\eta^2 = .034$. Graphs of the mean psychopathology ratings for the internalizing and externalizing items are presented in Figure 5.

Figure 5

Means of psychopathology measures (CBCL-DOF) for both internalizing and externalizing items

The analyses for the effects of Method and Referral on these ratings of externalizing and internalizing psychopathology were conducted with univariate ANOVAs. The externalizing ratings were not significant for Referral $F(2, 228) = 2.01$, $p > .05$, nor by Method $F < 1$, and not by their interaction, $F > 1$.

The internalizing ratings were significantly affected by Referral, $F(2, 228) = 4.53$, $p < .05$; but not by Method, $F < 1$; nor by their interaction, $F < 1$. The relationship between referral and the internalizing psychopathology ratings were mild to moderate, partial $\eta^2 = .034$. The externalizing ratings were neither significantly affected by Method, $F < 1$; Referral, $F(2, 228) = 2.01$, $p > .05$; nor their interaction, $F(6, 228) = 1.43$, $p > .05$. The source tables for these internalizing and externalizing psychopathology ratings are presented in Table 13.

Table 13

Univariate source tables for the effect of Referral and Method on the internalizing and externalizing psychopathology measures

Dependent Variable = internalizing psychopathology

Source	df	SS	MS	F	p
Method	3	142.03	47.34	.58	.63
Referral	2	740.21	370.12	4.53	.01
Method x Ref	6	190.36	31.72	.39	.89
Error	228	18619.99	81.67		

Dependent Variable = externalizing psychopathology

Source	df	SS	MS	F	p
Method	3	163.66	54.55	.40	.75
Referral	2	549.17	274.58	2.01	.14
Method x Ref	6	117.00	19.50	.14	.99
Error	228	31136.74	136.57		

A post-hoc analysis of the internalizing psychopathology measures revealed that participants who were given internalizing referrals tended to rate

internalizing behavior higher than those participants who were given general psychopathology referrals, Tukey a , $M_{diff} = 4.51$, $\alpha_{FW} < .05$. This analysis is presented in Table 14.

Table 14

Mean differences between referral groups on internalizing psychopathology measures

Referral	INT	EXT	GP
INT	--	3.23	4.51**
EXT		--	1.28
GP			--
** $p < .01$			

Discussion

This study sought to answer three questions regarding the utility of computer observation methods of data recording in relation to other methods: (a) do computer observation methods facilitate better accuracy in recording than other methods while in the presence of potential referral bias?, (b) are recorders who use computer observation methods more confident in their estimates than recorders who use other methods?, and (c) are recorders who use computer observation methods less influenced by confirmatory bias than recorders who use other methods?

Effects on Accuracy

To answer the first question, the overall analysis found no interaction between Method and Referral on observational accuracy; however, there were significant main effects detected for both of the independent variables.

Although the hypothesized interaction between Method and Referral on the combined accuracy measures was not observed in the overall analysis, an exploratory analysis showed an interaction when participants measured "verbal disagreement" on the externalizing

domain. This interaction presented some evidence that computer observation methods may facilitate greater accuracy in the presence of referral bias than other methods. This was evident in that the variability in accuracy among the referral groups increased respectively among the partial-interval, narrative, and no-method group conditions.

The presence of an interaction on just one of four targets may have occurred because it was more difficult to discriminate "verbal disagreement" from the other targets. Specifically, the target definition for a verbal disagreement incorporated three criteria: the video-tape actor stating (a) "I won't do this", (b) "I don't want to do this", or (c) "I want to do (something else)". The other target definitions included just one criterion (e.g., "record when the videotape actor states 'I'm stupid'"). When a verbal disagreement was suspected, each participant had to check each behavior criterion in order to confirm its occurrence. In this situation, participants in the narrative and partial-interval groups were additionally occupied with the mental and physical demands of recording this behavior whereas participants in the computer observation group

simply had to click a button to log an observation. Thus, participants in the computer observation group were able to evaluate the occurrences of "verbal disagreement" more accurately than other groups.

This variability in accuracy is reflected in the kappa-coefficient agreement ratings for each target criterion that was established by the five graduate students. The verbal disagreement target had the lowest agreement rating (i.e., $\kappa = .89$), suggesting that more variability would be observed when estimating behavior occurrences for this target.

The analysis of the main effect for Method on accuracy revealed that interval-recording methods (i.e., computer observation and partial-interval form) facilitated greater observational accuracy than narrative or "no-method" formats. This effect was strong (partial $\eta^2 = .18$), and was evident for observational accuracy on both internalizing (i.e., self-labeling as "stupid"; stomach complaints) and externalizing (i.e., verbal disagreement; switching activities) dimensions. Thus, computerized and partial-interval observational methods may yield higher levels of accuracy than

narrative or no-method techniques for observing discrete, moderate-frequency target behaviors like those that were presented in this study.

Overall, participants who used the computer observation method produced more accurate observational data than two other method groups, and equivalent accuracy for the third partial-interval group. There was also evidence that participants in the narrative group produced more accurate estimates than participants who received no training (i.e., the no-method group) only when measuring "verbal disagreement." This outcome was possibly due to a combination of the additional training the narrative group received over the no-method group and the difficulty in discriminating "verbal disagreement" from the other targets.

The effect of Referral on observational accuracy presented two surprising results. First, the effect of Referral was only significant when the participants measured "verbal disagreement." It was expected that the internalizing and externalizing behavior estimates would be inflated with respect to each referral condition, and thus would be less accurate than those estimates from the general psychopathology referral

group. This outcome was expected because participants may have selectively attended to behaviors that supported the initial referral (Arkes & Harkness, 1980; O'Reilly et. al, 1989). Conversely, the effect of the Referral information on accuracy may have served to heighten the participants' awareness to the target definitions and occurrences, thus resulting in the mostly non-significant effects for Referral on observational accuracy.

In addition, the referrals in this study were presented with a goal for accuracy and probably did not provide as strong of an incentive for confirmatory bias as might be found in applied settings. Therefore, it is possible that pervasive main effects for Referral were not seen in this study because the motivation to conform with the referral was minimal. Subsequent research on the effects of referral information on observation accuracy might include the presence of deadlines, conditional rewards, or self-disclosure of results to an expert to simulate referral pressure in applied settings.

The second surprising result was that the externalizing referral group produced more accurate

estimates on the target "verbal disagreement" than the participants in other referral groups. This unexpected outcome may have resulted because the participants were provided with referral conditions before the target definitions were read. Therefore, the participants in the externalizing referral condition may have been more attentive to the externalizing target definitions than the participants from the other referral conditions. It was previously mentioned that the externalizing-target definitions comprised more parameters than the internalizing targets. Therefore, hyper-awareness to the externalizing target definitions possibly contributed to the accuracy of the externalizing referral group.

Effects on Confidence

The second question this study addressed participants' levels of confidence in their ratings. The majority of the analyses indicate that participants using computer observation methods were more confident in their estimates than participants using other methods. As predicted, the participants using computerized methods rated their confidence higher than participants using the other methods; however, this was

only the case when participants recorded "switching" behavior. When measuring "verbal disagreement," the computer observation group evidenced more confidence than the partial-interval group, but not more than the narrative and no-method groups.

When measuring the internalizing behaviors, participants in the computer observation and partial-interval groups were more confident in their estimates than participants in the narrative group. Additionally, the computer observation group demonstrated more confidence in their estimates than the no-method group; however, the partial-interval group did not report higher confidence levels than the no-method group.

The analysis of observer confidence was conducted as an exploratory analysis to provide directions for future research. Specifically, previous studies (Arkes, 1980; Einhorn & Hogarth, 1978; Faust et. al, 1988) have found that individuals who are overconfident in their data generally adhere to hypotheses that are founded on these data. Therefore, it is possible that computer observation, and possibly partial-interval, recording methods might be at risk for confirmatory bias due to overconfidence. It is also possible, however; that the

confidence reported by the computer observation and partial-interval participants was appropriate and not conducive to confirmatory bias. The relationship between confidence, observation techniques, and decision-making practice will need to be explored in future research.

Effects on Measures of Psychopathology

The third research question explored the potential effects that referral information and observational methods would have on biasing the clinical interpretations of the observer. The clinical interpretations were measured by a global scale, which was constructed by the experimenter, and the Achenbach Direct Observation Form (CBCL-DOF; Achenbach, 1986). The global scales were used as a validity check for the clinical measure, as some of the items on the CBCL-DOF closely matched the content in the instructions that were read to the participants. The data measured by both instruments produced similar results, however, suggesting that the participants' interpretation was not biased by the language of each instrument and the wording of the target definitions. On both instruments, the observers' interpretations were more easily biased

by internalizing-content referrals than by general psychopathology-content referrals.

As predicted, the participants who were given internalizing information tended to rate internalizing symptomatology higher than participants who were given general psychopathology information. An unexpected finding was that no significant differences were found among the Referral or Method groups in the reports of externalizing symptomatology.

Interestingly, participants who received the internalizing referral rated externalizing symptomatology similarly to the other groups. This outcome was unexpected given the literature on referral information contributing to confirmatory bias (Arkes & Harkness, 1980; O'Reilly et. al, 1989). A possible explanation for this is that participants who received the internalizing referral were not as prepared to observe highly salient externalizing behaviors as participants receiving the externalizing referral who would have expected these behaviors. Previous research has demonstrated that externalizing behaviors are more often a cause for concern and referral than internalizing behaviors among school teachers (Garland,

1995; Skiba, Peterson, & Williams, 1997). As the internalizing-referral participants watched the videotape, they may have correctly wondered why oppositionality and hyperactivity were not mentioned in the referral. Thus, these participants may have remembered and over-reported externalizing behaviors because they were unexpected and caused concern. The participants who were not expecting to see an abundance of internalizing behaviors did not over-report the internalizing targets, perhaps because these targets were not perceived as great of a concern as the externalizing targets.

Limitations of the Research

This study presents some limitations that must be considered when reviewing the results for interpretation. While a key advantage of this study was that it was conducted in a controlled laboratory setting that controlled for threats of internal validity, the disadvantages concern the interpretation of the causal agent (i.e., construct validity), and the application of these results to other settings (i.e., external validity). The issues concerned with the internal,

construct, and external validity of this study will now be addressed.

Limitations related to internal validity. The primary limitation related to the internal validity of this study concerned the establishment of the criterion for accuracy. The criterion was established by having graduate students observe the videotape using a partial-interval recording method. It is possible that the criterion may have given the partial-interval and computer observation groups an unfair advantage over the narrative and no-method groups by using the method upon which the criterion was established. The formatting of the no-method and narrative methods, being radically different from the time-formatted partial-interval and computer observation methods, were likely to produce different data. Had the criterion been established to "identify and estimate targets for a functional behavioral analysis," perhaps the narrative and no-method groups would have produced more useful information than the partial-interval or computer observation method groups. The implication here is not that narrative and no-method techniques are of limited use; it is that narrative and no-method techniques do

not lend themselves to accurate recording when clearly defined targets of moderate frequency are being recorded.

This study was otherwise conducted so that other common threats to internal validity (i.e., history, maturation, regression to the mean, selection bias, instrumentation, and attrition; Kazdin, 2000) were addressed. Specifically, major threats to internal validity were controlled for via random selection and assignment of participants to groups, and via having the study conducted in a controlled and scripted environment.

Limitations related to construct validity. The limitations that may have affected the construct validity of this study focused on the possibility of the design to detect real effects in accuracy between the partial-interval and computer observation groups. The construct validity of a study has to do with determining the basis of the causal relationship that was demonstrated, assuming that threats to internal validity have been satisfactorily accounted for (Kazdin, 1998). The true effect of computer observation methods offering an improvement in accuracy over partial-interval methods

might be better detected by including more criteria for defining their targets, by including more targets to observe, or by extending the observational time period.

It is possible that the differences in accuracy between the computer observation and partial-interval groups might not have been detected because there were a limited number of concretely defined targets, thus potentially equalizing the cognitive load of the observation tasks for both groups. The targets for this study were selected and defined in a manner that would present equal representation of internalizing and externalizing behaviors. These adjustments resulted in having participants only observe four targets. It is speculated that a stronger effect for Method would have been seen had the participants tracked more than four target behaviors.

In addition, three of the four target behaviors used in this study were defined in a fairly direct and clear manner. Only the "verbal disagreement" target included more criteria than the other targets, and thus presented a greater chance for error in the measurement of this target. With this chance for error being, an interaction effect of Method and Referral on accuracy

for measuring "verbal disagreement" was seen. Thus, evidence for an effect for Method when the target definition was more complex.

Lastly, the short-term observational duration in this study made it relatively immune to the effects of observer drift, which perhaps would have better differentiated between the computer observation and partial-interval groups (Skinner et. al, 2000).

Limitations related to external validity. The most apparent limitations of this study are related to external validity. The following discussion related to external validity will focus first on how the specificity of the experimental condition limits the broad application of these results to other setting. Then the focus will turn to how the selection and training of the participants also limits the external validity of this study.

Behavioral observation techniques are highly diverse; the methods, targets, and function of an assessment vary according to the purpose and setting of the assessment. The arrangement of this study, being focused on assessing interval-recording methods in a controlled setting, may have limited its application to

measuring behaviors in other situations. The following discussion will specifically address how the interval-method focus, the format for the observations, the group recording environment, the limited subject reactivity, and the nature of the referral may have limited the external validity of this study.

First, this study was constructed to incorporate a stimulus environment that would be appropriate for an interval or otherwise time-based observational procedure. Specifically, the actor was instructed to emit discreet target behaviors at a moderate frequency for a 10-minute observation. It can certainly be the case that other methods would be more appropriate for assessing different behaviors because they are too infrequent, too frequent, or the behaviors do not have a clear beginning or ending (e.g., crying). In addition, an observer might be more interested in establishing measures of the duration or intensity of a behavior. In these instances, interval-recording methods might not be able to yield data that are as useful as other observational methods (e.g., event recording, time-sampling, duration, or latency recording). Therefore, the results of this study may be safely generalized only

to the performance of observers who use partial-interval and computer observation methods of recording.

Second, the study constructed the elements of the observation that normally would introduce a potential source of bias. Specifically, it was not necessary for the participants to select, arrange, and clearly define the targets. Additionally, the setting of the observation was already selected for the participants, thus eliminating the need to question whether this sample of behavior was a representative one and to observe other settings where the behaviors might fluctuate.

Third, the participants recorded their observations in a group. This presented a potential opportunity for the participants to compare their own recordings with the other participants and adjust their recordings if they were not similar to the others. This may have enhanced the accuracy of the participants in this study when compared to observers in applied settings because it presented a greater opportunity to establish interrater reliability.

Fourth, the observers watched a video-tape of an actor who was performing in response to cues. This led

to a presentation that is qualitatively different than a child experiencing true behavioral difficulties in a natural setting. In this situation, the common effects of subject reactivity, peer interactions with the observer, and classroom distractions (e.g., students greeting the observer), were not factors that detracted from accuracy. In addition, the effect of observer drift was minimal because the participants were only required to observe 10 minutes of videotape.

Last, and perhaps most importantly, the referrals in this study were presented with a goal for accuracy with no conditions attached. The referrals in applied settings may, directly or inadvertently, be presented in a way that sets an incentive to find behaviors that are in concordance with the referral sources' beliefs or desires (Arkes & Harkness, 1980; Bargh & Chartrand, 1999). Additionally, the participants in this study only had two sources of information to consider in their decisions regarding psychopathology. It is often the case that clinicians are confronted with many sources of data, and diagnostic accuracy would be reduced as more sources of data are combined in the decision-making process (Faust & Nurcombe, 1988; Grove et. al, 2000).

These factors may help to illuminate why the participants in this study demonstrated lower-than-predicted referral bias, and why clinicians in applied settings might be vulnerable to more sources of bias affecting the accuracy of their decisions.

The selection of the participants in this study was limited to undergraduates taking introductory level classes who had no prior training in observational methods. This is an area of concern because the sample was not screened for several variables that may have influenced the results. For example, previous research has identified factors such as SES (Alevizoa, DeRisi, Liberman, Eckman, & Callahan, 1978), high verbal and clerical skills (Skindrud, 1973), motivation (Dancer et. al, 1978), and morale (Guttman, Spector, Sigal, Rakoff, & Epstein, 1971) that can affect observational accuracy. These traits might be considered when conducting this or a similar observational study that selected participants from another sample pool. For example, recruiting highly trained observers that are familiar with one particular observation method may produce different results than this study.

Lastly, the training of the participants also may have imposed limitations on external validity in this study. Specifically, participants in the narrative, partial-interval and computer observation groups underwent a short (i.e., 15-minute) training period. The concern here is that intense training has been found to cause higher measures of agreement and accuracy in observations (Foster & Cone, 1986). The results of this study, however, suggest that the training in the methods, particularly the partial-interval and computer observation methods, were adequate to produce acceptably accurate observations from participants with little previous experience in observation. The effects of the limited training in this study were minor and, therefore, these results can be reasonably generalized to clinical settings.

Implications of the Research

The outcomes of this research support the use of computer observation and partial-interval recording methods as part of a best practice for obtaining data on moderate to high rates of discreet behaviors. This study has also established respectable validity in using relatively new computer observation methods in the place

of partial-interval methods to collect interval data. This information is of use to researchers and clinicians who are considering using computer observation techniques in their work. Specifically, some researchers and practitioners may be in a position to decide whether they want to invest in purchasing computer observation software or to use a partial-interval checklist for tracking behavior. It is important to mention here that computer observation methods also reduce the time-consuming and error-prone practice of hand scoring and graphing observational data. In addition, computer observational data would be able to be stored in smaller spaces and be organized and protected in a computer file system (Kahng & Iwata, 2000; Sandman et. al, 2000). Depending on the need and frequency of observational assessments, the purchase of computer observation software could be a sound investment given the results of this study.

An additional best-practice implication is that narrative formats and unstructured observations are best used for collecting qualitative information for hypothesis generation. These methods are not well suited to collecting systematic, time-specific trend

data because they tend to facilitate overestimation of problem-behavior occurrences.

Finally, it is important to note that internalizing referrals brought about over-reporting of externalizing behavior in all of the method groups. It is hypothesized that this occurred because the participants became inattentive to the externalizing target definitions that followed the referral, and they were unprepared to see the externalizing behaviors. Practice implications suggest that incomplete or inaccurate referral information may be associated with an oversensitivity to particular behaviors.

Directions for Future Research

This study presents a foundation for validating computer observation and partial-interval methods in their recording accuracy, and their potential to facilitate accurate diagnostic decisions. Additional evidence was found suggesting that computerized observation methods can enhance the accuracy of an observation when certain variables are present. Specifically, computerized methods seem to improve upon accuracy when multiple criteria are used to define the targets. Therefore, a study that compares computerized

methods to other interval-recording methods with variations on the complexity and number of the targets might clarify accuracy differences between computerized and partial-interval methods. Additional studies might also explore the effects observing more targets or high-frequency targets to further define accuracy differences between methods.

A possible implication relating observer confidence, observation method, and decision-making accuracy was not clear given the results of this study. On one hand, there is a potential downside of using computer observation methods in that they might lead to overconfidence in the data. This overconfidence can lead to over-reliance on original hypotheses that are often generated by referral information (Arkes & Harkness, 1980; Einhorn & Hogarth, 1978). On the other hand, it is possible that the people using computer observation methods are appropriately confident in their data and this mind-set would not lead to over-reliance on original hypotheses. The relationship between observer confidence, observation method, and decision-making accuracy must be explored in future research.

In addition, a study of the manner in which clinicians combine multiple data sources with direct behavior observation data to reach decisions would be valuable to applied science. The finding of overconfidence in computer observation methods might serve as a starting point for this analysis. A follow-up study of this finding might introduce the results from another behavioral assessment source (e.g., an observation or parent/teacher rating scale) in addition to the observation. Here, measures on both the confidence in the observation data and confidence in the diagnostic decision could be obtained and analyzed.

Finally, it would be useful to study the effects of computerized observation methods with real subjects, professional observers, and longer observation durations to see if the results of this present study would generalize to applied settings.

Summary

This study presents support for computerized and partial-interval techniques as methods that enhance greater observational accuracy than narrative and unstructured recording methods. An exploratory analysis presented evidence that observers using computerized

methods might be able to record multiple-criteria targets with greater accuracy than partial-interval methods. This relationship, however, needs to be studied further to make the claim that computerized methods facilitate greater accuracy than partial-interval methods. Additional findings of the study were that internalizing referrals induced over-reporting of externalizing behavior and higher ratings of internalizing psychopathology. Also, participants using computerized methods were more confident in their data than participants using other methods. These findings apply to research and clinical settings that are considering computerized methods to record interval data and integrating these data into clinical decisions. Finally, it was suggested that exploring the potential differences between computerized and partial-interval methods using more complex targets, samples of practitioners, and simulated referral conditions would be a logical direction for future research.

Diana V. Brown
_____
Diana V. Brown
Director of Compliance

_____
Date

Appendix B

Title (10 points) Child Observation Study

RESEARCHER

**Charles Sicotte, Graduate Student in School Psychology; Mcsicotte@aol.com
Phone: 277-5165**

CRITERIA FOR PARTICIPATION University student

TIME REQUIRED FOR PARTICIPATION

30-40 minutes

BRIEF DESCRIPTION

In this study you will first observe a 15-minute videotape of a child, and then you will be asked to respond to questions regarding what you saw.

The purpose of this study is to investigate how psychologists collect information through observation. The results will be useful for psychologists and teachers who use frequently use observation as a tool to collect information on their clients.

How to participate:

If you are interested in participating, the videotape will be shown in Quinn 217 at 2:00 every Tuesday and Thursday starting 3/7/01. Please sign up for the date and time you want to participate. Sign-up sheets are posted in the Psychology Department Office, 10 Woodward Hall (on the wall to the left of the Department Door).

Appendix C

Child Observation Study Sign-up Sheet

Please call Charles Sicotte at 277-5165 for more details

Tuesday; April 23, 2002 11:00am
Quinn Hall 217

Name

Phone

- 1) _____
- 2) _____
- 3) _____
- 4) _____
- 5) _____
- 6) _____
- 7) _____
- 8) _____
- 9) _____
- 10) _____
- 11) _____
- 12) _____
- 13) _____
- 14) _____
- 15) _____

In this study you will be observing a videotape of a child and will be asked questions about it. The study should take between 30-40 minutes to complete. Thank you for your interest.
—Charles Sicotte; phone 277-5165.

Appendix D

TO: PARTICIPANTS IN THE CHILD OBSERVATION STUDY

FROM: CHARLES SICOTTE, EXAMINER

Thank you for your interest in the Child Observation Study. The purpose of this study is to investigate the way people who work with children collect information through observation. The results will be useful for psychologists and teachers who use frequently use observation as a tool to collect information on their clients.

In the first part of this study you will be asked to observe a child in a videotape for 15 minutes, and then you will be asked to fill out forms regarding what you saw for the next 15 minutes. I cannot tell you exactly what you will see and do today because this might affect the results. However, I can let you know that the observation you will be doing will be of minimal risk to you and anyone involved in this study. Your participation in this study is completely anonymous. At the end of our session today you will be informed about the independent and dependent variables that were investigated for your PIA credit. The results of this study cannot be given until all data have been collected, though I will be happy to share the results of this study by posting them in Woodward Hall when the data analysis is complete.

You must be at least 18 years old to be in this study. It is important to understand that you reserve the right to end your participation at any time you wish. It is also important that your presence for the duration of the study is necessary for your participation to be factored in the results. Your full participation may be used for partial credit for any psychology courses granting credit for research participation.

I sincerely thank each of you for your participation, for making this project possible, and for potentially improving the lives of children in our schools. If you have any more questions or concerns about this study, you may contact the University of Rhode Island's Vice Provost for Graduate Studies, Research and Outreach, 70 Lower College Road, Suite 2, URI, Kingston, RI, (401) 874-4328.

Sincerely, Charles Sicotte, M.A.

Phone: 401-277-5165

Student Signature: _____ Date: _____

Examiner Signature: _____

Appendix E

CHILD OBSERVATION STUDY – PARTICIPATION FORM

PARTICIPATION CODE: _____

DATE: _____

DATE OF BIRTH: _____

CLASS: FRESHMAN   SOPHOMORE   JUNIOR   SENIOR   GRADUATE

MAJOR: _____

WILL YOU RECEIVE ACADEMIC CREDIT FOR PARTICIPATING IN THIS STUDY?   YES
NOIF YOU ARE RECEIVING ACADEMIC CREDIT FOR PARTICIPATING IN THIS STUDY, FOR
WHICH CLASS WILL YOU RECEIVE CREDIT?

DO YOU WEAR PRESCRIPTION GLASSES OR HEARING AIDS?   YES   NOIF YES, DO YOU HAVE THEM WITH YOU TODAY?   YES   NO

Appendix F

Participant Code: _____

Date: _____

Observation 1

1 Minute   Sec   **Disagre**   **Stupid**   **Switch**   **Stomach**

0-10				
11-20				
21-30				
31-40				
41-50				
51-60				

2 Minute   Sec   **Disagre**   **Stupid**   **Switch**   **Stomach**

	0-10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

Appendix G

Participant Code: _____

Date: _____

Observation 4

	Sec	Disagre	Stupid	Switch	Stomach
1 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
2 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
3 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
4 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
5 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
6 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
7 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
8 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
9 Minute	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

	Sec	Disagre	Stupid	Switch	Stomach
10 Min.	0 - 10				
	11-20				
	21-30				
	31-40				
	41-50				
	51-60				

Appendix H

NARRATIVE RECORDING FORM

CODE ON PACKET: _____***PLEASE USE THIS FORM TO RECORD WHAT YOU SEE IN THE VIDEOS**

Appendix I

CHILD OBSERVATION STUDY – RATING FORM

PARTICIPANT CODE: _____

DATE: _____

- 1) PLEASE INDICATE THE PERCENTAGE OF THE TIME YOU FELT THAT THE CHILD ENGAGED IN EACH OF THE BEHAVIORS. YOU MUST WRITE A PERCENTAGE.
- 2) PLEASE RATE YOUR CONFIDENCE IN THIS ESTIMATE BY CIRCLING YOUR CHOICE ON THE 7-POINT SCALE BELOW EACH ITEM.

WHAT PERCENTAGE OF THE TIME DO YOU THINK THE CHILD DISAGREED WITH HER MOTHER?

_____ %

Please indicate your confidence in this percentage:

1	2	3	4	5	6	7
No Confidence	Minimal Confidence	Some Confidence	Acceptable Confidence	Very Confident	Strongly Confident	Absolutely Certain

WHAT PERCENTAGE OF THE TIME DO YOU THINK THE CHILD CALLED HERSELF “STUPID”?

_____ %

Please indicate your confidence in this percentage:

1	2	3	4	5	6	7
No Confidence	Minimal Confidence	Some Confidence	Acceptable Confidence	Very Confident	Strongly Confident	Absolutel Certain

WHAT PERCENTAGE OF THE TIME DO YOU THINK THE CHILD SWITCHED FROM HER HOMEWORK TO ANOTHER ACTIVITY?

_____ %

Please indicate your confidence in this percentage:

1	2	3	4	5	6	7
No Confidence	Minimal Confidence	Some Confidence	Acceptable Confidence	Very Confident	Strongly Confident	Absolutely Certain

WHAT PERCENTAGE OF THE TIME DO YOU THINK THE CHILD COMPLAINED ABOUT A STOMACH ACHE?

_____ %

Please indicate your confidence in this percentage:

1	2	3	4	5	6	7
No Confidence	Minimal Confidence	Some Confidence	Acceptable Confidence	Very Confident	Strongly Confident	Absolutely Certain

Appendix J

CHILD OBSERVATION STUDY – GLOBAL RATING FORM

PARTICIPANT CODE: _____

DATE: _____

PLEASE INDICATE HOW MUCH YOU FELT THE CHILD DEMONSTRATED THE FOLLOWING BEHAVIORS BY CIRCLING THE MOST APPROPRIATE NUMBER:

ONE WAY TO CLASSIFY CHILD BEHAVIOR PROBLEMS VERY GENERALLY IS AS “INTERNALIZING.” THIS IS A GROUP OF BEHAVIOR PROBLEMS SHOWING SYMPTOMS OF ANXIETY, DEPRESSION, AND WITHDRAWAL.

HOW MUCH DID THE CHILD IN THE VIDEO DEMONSTRATE INTERNALIZING BEHAVIORS?

1	2	3	4	5	6	7
No Internalizing	Minimally Internalizing	Some Internalizing	Moderately Internalizing	Highly Internalizing	Very Highly Internalizing	Extremely Internalizing

ANOTHER WAY TO CLASSIFY CHILD BEHAVIOR PROBLEMS VERY GENERALLY IS AS “EXTERNALIZING.” THIS IS A GROUP OF BEHAVIOR PROBLEMS SHOWING SYMPTOMS OF HYPERACTIVITY, DISOBEDIENCE, AND AGGRESSION.

HOW MUCH DID THE CHILD IN THE VIDEO DEMONSTRATE EXTERNALIZING BEHAVIORS?

1	2	3	4	5	6	7
No Externalizing	Minimally Externalizing	Some Externalizing	Moderately Externalizing	Highly Externalizing	Very Highly Externalizing	Extremely Externalizing

HOW MUCH ANXIETY DID THE CHILD IN THE VIDEO DEMONSTRATE?

1	2	3	4	5	6	7
No Anxiety	Minimal Anxiety	Some Anxiety	Moderate Anxiety	High Anxiety	Very High Anxiety	Extreme Anxiety

HOW HYPERACTIVE WAS THE CHILD IN THE VIDEO?

1	2	3	4	5	6	7
Not Hyperactive	Minimally Hyperactive	Some Hyperactivity	Moderately Hyperactive	Highly Hyperactive	Very Highly Hyperactive	Extremely Hyperactive

HOW DEPRESSED WAS THE CHILD IN THE VIDEO?

1	2	3	4	5	6	7
Not Depressed	Minimally Depressed	Some Depression	Moderately Depressed	Highly Depressed	Very Highly Depressed	Extremely Depressed

HOW DISOBEDIENT WAS THE CHILD IN THE VIDEO?

1	2	3	4	5	6	7
Not Disobedient	Minimally Disobedient	Some Disobedience	Moderately Disobedient	Highly Disobedient	Very Highly Disobedient	Extremely Disobedient

HOW WITHDRAWN (E.G., NOT SOCIAL) WAS THE CHILD IN THE VIDEO?

1	2	3	4	5	6	7
Not Withdrawn	Minimally Withdrawn	Some Withdrawal	Moderately Withdrawn	Highly Withdrawn	Very Highly Withdrawn	Extremely Withdrawn

HOW AGGRESSIVE WAS THE CHILD IN THE VIDEO?

1	2	3	4	5	6	7
Not Aggressive	Minimally Aggressive	Some Aggression	Moderately Aggressive	Highly Aggressive	Very Highly Aggressive	Extremely Aggressive

Appendix K

Participant

File Edit Help

Participant Code*

Date

Observation Time ☐ 2 Minutes ☒ 10 Minutes

Trial Number ☒ 1 ☐ 2 ☐ 3 ☐ 4

Target Behavior

0	0	0	0
Disobey	I'm Stupid	Switches	Stomach

Time of Observation

Test Time Sec

Interval Time

c:\windows\desktop Total # 0 10/24/20 3:24 PM

Appendix L

Referral Conditions

(a) Internalizing condition instructions - "...This child has been referred to the school psychologist because her teacher is concerned about her behavior. Specifically, this girl has demonstrated anxiety and depression, and has complained about body aches and pains. The purpose of your observation is to gain information that will be used in planning an intervention for this child that can be used by the teachers in her school."

(b) Externalizing condition instructions - "This child has been referred to the school psychologist because her teacher is concerned about her behavior. Specifically, this girl has demonstrated difficulty with attention and following directions, and has also demonstrated aggression. The purpose of your observation is to gain information that will be used in planning an intervention for this child that can be used by the teachers in her school."

(c) Neutral Condition instructions - "This child has been referred to the school psychologist because her

teacher is concerned about her behavior. The purpose of your observation is to gain information that will be used in planning an intervention for this child that can be used by the teachers in her school."

Appendix M

Training Sequences for Each Group

Instructions for No-method group

Good (morning) afternoon, I thank you for taking the time to be here today for this child observation study. In a few moments you will watch four short video clips of a child. I first want to direct your attention to what's inside your packet. If you have not opened it you may open it now. Inside I want you to find two informed consent forms that look like this, does everyone have these? Good, There are two of them; one for me to keep and one for you to keep. (Let me summarize the letter for you.) Signing this form will allow you to participate in this study, are there any questions? OK, please go ahead and sign it.

I would like you to find the form labeled "participation form" in your packet and fill it out; it looks like this (hold it up). This information will let me know more about you. Because your name is not required on this form this information will be confidential. The top portion of the form asks for your participation code. The first two spaces of this code is number and letter on your packet. For the third

space I would like you to fill in the number of the computer that you are sitting at. This number is on the upper right hand corner of the computer. After you have filled in your participant code, Go ahead and fill out the rest of the form (pause). If everyone is finished you may place the informed-consent and participation forms back inside the packet. Remember to keep one informed-consent form for yourself.

In a moment we'll be ready to watch the video clips of this child. But first let me tell you about this child and what you will be looking for in the videos. This is a 9-year old girl who attends a local school **[insert biasing conditions here]**.

Psychologists and teachers use observations as a way to help diagnose and intervene with children who are referred for services. So when you are doing your observations, I would like you to take the perspective of a professional who will eventually diagnose and design a treatment for this child.

The videos show the girl doing homework with her mother at home. I want you to look for 4 things the girl does in these videos. Listen carefully, they are:

verbally disagreeing with her mother, calling herself "stupid", switching activities, and complaining of stomach problems. Let me explain these behaviors in a little more detail. Listen carefully to these definitions because I will only explain them here. A verbal disagreement is any time the girl in the video says "no" to her mother, says that she does not want to do her work, or says that she wants to do something other than her work. Calling herself stupid means anytime the girl says "I'm Stupid", "I'm so stupid", "I don't know" or "I can't do this". The girl switches activities any time she gets up out of her seat, plays with items in the basket, plays with the dog, or plays with the soda can. Finally, "Stomach Problems" is defined as when the subject says "my stomach hurts" or "I need medicine".

The first three videos are 2-minutes long. During the first video I will identify these behaviors for you, in the second I'll mention the behaviors following the video. I'll give no feedback following the third video, this will be for practice. The final video will be 10-minutes long, and it will be where you get most of your information on this child. You will not be

allowed to write during your observations, simply watch the videos. Are we ready to begin?

Good. Here is the first video. (During the video; see here, this is a verbal disagreement; this is a switch, she went up to get the Coke can; see, she complain of her stomach; did you catch that, she called herself stupid.)

Instructions for Narrative group

Good (morning) afternoon, I thank you for taking the time to be here today for this child observation study. In a few moments you will watch four short video clips of a child. I first want to direct your attention to what's inside your packet. If you have not opened it you may open it now. Inside I want you to find two informed consent forms that look like this, does everyone have these? Good, There are two of them; one for me to keep and one for you to keep. (Let me summarize the letter for you.) Signing this form will allow you to participate in this study, are there any questions? OK, please go ahead and sign it.

I would like you to find the form labeled "participation form" in your packet and fill it out; it looks like this (hold it up). This information will let

me know more about you. Because your name is not required on this form this information will be confidential. The top portion of the form asks for your participation code. The first two spaces of this code are the number and letter on your packet. For the third space I would like you to fill in the number of the computer that you are sitting at. This number is on the upper right hand corner of the computer. After you have filled in your participant code, Go ahead and fill out the rest of the form (pause). If everyone is finished you may place the informed-consent and participation forms back inside the packet. Remember to keep one informed-consent form for yourself.

In a moment we'll be ready to watch the video clips of this child. But first let me tell you about this child and what you will be looking for in the videos, and how you will be keeping track of what you see. This is a 9-year old girl who attends a local school **[insert biasing conditions here]**.

Psychologists and teachers use observations as a way to help diagnose and intervene with children who are

referred for services. So when you are doing your observations, I would like you to take the perspective of a professional who will eventually diagnose and design a treatment for this child.

The videos show the girl doing homework with her mother at home. I want you to look for 4 things the girl does in these videos. Listen carefully, they are: verbally disagreeing with her mother, calling herself "stupid", switching activities, and complaining of stomach problems. Let me explain these behaviors in a little more detail. Listen carefully to these definitions because I will only explain them here. A verbal disagreement is any time the girl in the video says "no" to her mother, says that she does not want to do her work, or says that she wants to do something other than her work. Calling herself stupid means anytime the girl says "I'm Stupid", "I'm so stupid", "I don't know" or "I can't do this". The girl switches activities any time she gets up out of her seat, plays with items in the basket, plays with the dog, or plays with the soda can. Finally, "Stomach Problems" is defined as when the subject says "my stomach hurts" or "I need medicine".

The way I want you to keep track of these behaviors is to use the form labeled "Narrative Recording Form" in your packet, please take this out now. When you see either of the behaviors that we have just discussed occur, I want you to write them down on this form.

I will show you how to do this. The first three videos are 2-minutes long. During the first video I will identify these behaviors for you and write them on this overhead projector to show you how the observation is to be done. In the second video I'll let you observe on your own and then I will show you my notes following the video. The third video will serve as a practice for you, I will give no feedback. The final video will be 10-minutes long, and it will be where you get most of your information on this child. You will be using the same form to write your observations for all of the videos, so if you run out of room you are welcome to use the other side. Are we ready to start?

Good. Here is the first video. (During the video; see here, this is a verbal disagreement; this is a switch, she went up to get the Coke can; see, she

complain of her stomach; did you catch that, she called herself stupid.)

OK, you see how I wrote each of the observations down? I'd like you to do this on your own for the next video. Following the video I'll show you what I wrote to check and see if you wrote down all of your observations. Are we ready to start? OK here it is.
(show video)

OK, here is what I wrote down. You should have at least written down that she disagreed once, switched once, called herself stupid once, and complained of her stomach once. Do all of you have this?

All right, we are ready to watch the third 2-minute video for practice. Just write down what you see, I won't give you any feedback.

OK, now we are ready to watch the 10-minute video where you will be getting most of your information. Please do your best to write down your observations, I won't give you any feedback on this observation. Are you ready? OK, here it is.

Instructions for Partial-Interval Form group

Good (morning) afternoon, I thank you for taking the time to be here today for this child observation

study. In a few moments you will watch four short video clips of a child. I first want to direct your attention to what's inside your packet. If you have not opened it you may open it now. Inside I want you to find two informed consent forms that look like this, does everyone have these? Good, There are two of them; one for me to keep and one for you to keep. (Let me summarize the letter for you.) Signing this form will allow you to participate in this study, are there any questions? OK, please go ahead and sign it.

I would like you to find the form labeled "participation form" in your packet and fill it out; it looks like this (hold it up). This information will let me know more about you. Because your name is not required on this form this information will be confidential. The top portion of the form asks for your participation code. The first two spaces of this code is number and a letter on your packet. For the third space I would like you to fill in the number of the computer that you are sitting at. This number is on the upper right hand corner of the computer. After you have filled in your participant code, Go ahead and fill out

the rest of the form (pause). If everyone is finished you may place the informed-consent and participation forms back inside the packet. Remember to keep one informed-consent form for yourself.

In a moment we'll be ready to watch the video clips of this child. But first let me tell you about this child and what you will be looking for in the videos, and how you will be keeping track of what you see. This is a 9-year old girl who attends a local school **[insert biasing conditions here]**.

Psychologists and teachers use observations as a way to help diagnose and intervene with children who are referred for services. So when you are doing your observations, I would like you to take the perspective of a professional who will eventually diagnose and design a treatment for this child.

The videos show the girl doing homework with her mother at home. I want you to look for 4 things the girl does in these videos. Listen carefully, they are: verbally disagreeing with her mother, calling herself "stupid", switching activities, and complaining of stomach problems. Let me explain these behaviors in a

little more detail. Listen carefully to these definitions because I will only explain them here. A verbal disagreement is any time the girl in the video says "no" to her mother, says that she does not want to do her work, or says that she wants to do something other than her work. Calling herself stupid means anytime the girl says "I'm Stupid", "I'm so stupid", "I don't know" or "I can't do this". The girl switches activities any time she gets up out of her seat, plays with items in the basket, plays with the dog, or plays with the soda can. Finally, "Stomach Problems" is defined as when the subject says "my stomach hurts" or "I need medicine".

The way I want you to keep track of these behaviors is to use the forms inside your packet that have grids on them. They look like this and are labeled "Observation 1, 2, 3, and 4". Please take out these four forms and place the form that is labeled "Observation 1" in front of you.

These forms are designed to keep track of behaviors that occur, and the time that they occurred. Because we need to keep time, I have made a timer available on your

computer screen. Is there anyone that does not have a timer on their computer? This is important (pause).

You will be recording the four behaviors that I have described and the times that they occur. As you can see, the four behaviors appear on the top of the grid. Verbal disagreement is labeled "disagree", calling herself "stupid" appears next, "switch" stands for switching activities, and "stomach" stands for "complaining of stomach problems." (show this on overhead)

Notice that the 2 grids are labeled "1-minute" and "2-minute". This form is formatted for a two-minute observation. Let's look at the first minute grid (show on overhead projector), on the column here (indicate) you can see spaces indicating 10-second intervals. You see that there are spaces that indicate 0-10, 11-20, and so forth; these indicate the seconds that have passed.

The way this works is that you place a check mark in the space that corresponds to the behavior and the time that it occurs. For example, if the child says "I'm stupid" 5 seconds into the video, place a check mark here (demonstrate). If the child says "my stomach

hurts" 30 seconds into the video, place a check mark here (demonstrate).

I will start the video in a moment and we will do a 2-minute observation using the sheet in front of you for practice. I will demonstrate on the overhead how this observation is to be recorded. We will be looking for all four of the behaviors I have described during this practice session. Are there any questions? Good. I will wait until the minute hand is on the 12 to start the video, I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

Stop. (explain data)

Now you are going to perform another 2-minute observation. This time I will not demonstrate it on the overhead, but I will give you feedback to check yourself when you are finished. Please do this on the form labeled "Observation 2". I will give you a moment to find this form. Again we will be looking for the same four behaviors. Are there any questions? Good. I will wait until the minute hand is on the 12 to start the video, I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

(give feedback by showing overhead and answering any questions)

OK, Now you are going to perform your last 2-minute practice observation. This time I will not demonstrate or give feedback, it is for you to practice the technique. Please do this on the form labeled "Observation 3". I will give you a moment to find this form. Again we will be looking for the same four behaviors. Are there any questions? Good. I will wait until the minute hand is on the 12 to start the video, I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

OK, Now we are ready to do our final observation using the final form labeled "Observation 4". This observation is 10-minutes long and it is where you will obtain the most information on this child. I will not demonstrate or give feedback for this observation; Please do the best that you can. Again we will be looking for the same four behaviors. Are there any questions? Good. I will wait until the minute hand is on the 12 to start the video, I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

Computer Observation Condition Training

Good (morning) afternoon, I thank you for taking the time to be here today for this child observation study. In a few moments you will watch four short video clips of a child. I first want to direct your attention to what's inside your packet. If you have not opened it you may open it now. Inside I want you to find two informed consent forms that look like this, does everyone have these? Good, There are two of them; one for me to keep and one for you to keep. (Let me summarize the letter for you.) Signing this form will allow you to participate in this study, are there any questions? OK, please go ahead and sign it.

I would like you to find the form labeled "participation form" in your packet and fill it out; it looks like this (hold it up). This information will let me know more about you. Because your name is not required on this form this information will be confidential. The top portion of the form asks for your participation code. The first two spaces of this code is number and letter on your packet. For the third space I would like you to fill in the number of the computer that you are sitting at. This number is on the upper

right hand corner of the computer. After you have filled in your participant code, Go ahead and fill out the rest of the form (pause). If everyone is finished you may place the informed-consent and participation forms back inside the packet. Remember to keep one informed-consent form for yourself.

In a moment we'll be ready to watch the video clips of this child. But first let me tell you about this child, what you will be looking for in the videos, and how you will be keeping track of what you see. This is a 9-year old girl who attends a local school **[insert biasing conditions here]**.

Psychologists and teachers use observations as a way to help diagnose and intervene with children who are referred for services. So when you are doing your observations, I would like you to take the perspective of a professional who will eventually diagnose and design a treatment for this child.

The videos show the girl doing homework with her mother at home. I want you to look for 4 things the girl does in these videos. Listen carefully, they are: verbally disagreeing with her mother, calling herself

"stupid", switching activities, and complaining of stomach problems. Let me explain these behaviors in a little more detail. Listen carefully to these definitions because I will only explain them here. A verbal disagreement is any time the girl in the video says "no" to her mother, says that she does not want to do her work, or says that she wants to do something other than her work. Calling herself stupid means anytime the girl says "I'm Stupid", "I'm so stupid", "I don't know" or "I can't do this". The girl switches activities any time she gets up out of her seat, plays with items in the basket, plays with the dog, or plays with the soda can. Finally, "Stomach Problems" is defined as when the subject says "my stomach hurts" or "I need medicine".

You will be using the computer program in front of you to keep track of the behaviors I have just described. Before I describe the program, I want you to enter the letter (A,B, or C) and the number of the computer you are sitting at in the upper left hand space (demonstrate). I also want you to make sure that your computer is selected for "2-minute observation" by having this dot highlighted (show).

On your screen you will notice a "start" button, and four buttons labeled "verbal disagreement", "I'm stupid", "switch", and "stomach" that correspond with behaviors I have defined. After pressing the start button, you will record behaviors by clicking the appropriate button once. For example, if you hear the child say "I'm stupid" you should click the "stupid" button once like this. If you see the child switch activities you should click the "switch" button once like this. Just click the appropriate button any time you see a behavior occur. Let me point out some other things that will be happening on the screen. When you click a button, the button will turn red for a few moments to indicate that you clicked it and then it will return to regular. There is an option to "undo" an mistaken entry by clicking this "Undo" button. Also, see these bars here, they indicate the time elapsed during the observation. Don't pay as much attention to these bars (point out), I want you to pay attention to the video.

The computer will keep count of all the behaviors you indicate and the time that they occurred. At the end of the observation it will display the percentage of

time the child engaged in the behavior. I will start the video in a moment and we will do a 2-minute observation using the computer for practice.

Please check to see that you entered the letter (A,B,C) and the number of your computer. If you have, your computer is ready to start the 2-minute observation. For this trial, I will be demonstrating the data entry on this screen. Remember, you will be looking for all four of the behaviors I have described during this practice session. Are there any questions? (questions will be addressed by repeating the criteria and/or saying "just do the best you can"). Good, I will count down from five to signal the start of the observation. When I say "go" I want you to click the start button and begin observing. Ready?; 5, 4, 3, 2, 1.Go.

Stop. (explain data) Your observation data has been saved.

Now you are going to perform another 2-minute observation. Your computer should already be formatted for a 2-minute observation by having this dot

highlighted (point) and the dot labeled "2" should be highlighted. This time I will not demonstrate, but I will give you feedback to check yourself when you are finished. Are there any questions? Good. When I say "go", I want you to press the "start" button and begin observing. I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

(give feedback by showing overhead of computer screen with percentages; answer any questions) Your observation has been saved.

OK, Now you are going to perform your last 2-minute practice observation. This time I will not demonstrate or give feedback, it is for you to practice the technique using the computer. Your computer should already be formatted for this observation by having the dot labeled "2-minute" highlighted (point) and the dot labeled "3" should be highlighted. Are there any questions? Good. When I say "go", I want you to press the "start" button and begin observing. I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

OK, Now we are ready to do our final observation. Your computer should have automatically been set to "10-

minute" and "4" under "trial". This observation is 10-minutes long and it is where you will obtain the most information on this child. I will not demonstrate or give feedback for this observation; Please do the best that you can. Again we will be looking for the same four behaviors. Are there any questions? Good. . When I say "go", I want you to press the "start" button and begin observing. I will count down from five to signal the start of the observation; 5, 4, 3, 2, 1.Go.

Appendix N

Post-Observation Instructions

PIF Group

Now that we are finished watching the videos, it's time to analyze our observation. I want you to add up all of the check marks for each behavior you observed. You can start by counting the number of times the girl in the video verbally disagreed with her mother during the 10-minute segment. Write this number underneath the "disagree" column and tally up the rest of the behaviors. When you are finished tallying all of the behaviors, write the percent engaged in the behavior using this chart as a reference.

Now you are ready to evaluate this child. Take out the form labeled "Rating form", this form asks you to estimate the percent engaged in each behavior and your confidence in that estimate. Please take a few minutes to fill out this form.

Next, I want you to fill out the "Global Rating Form" which asks your opinion on general aspects of this child's behavior. It looks like this (hold up) Please take a few minutes to fill out this form.

The final thing that you'll do today is to fill out a more detailed form asking specific questions about his child's behavior. The form looks like this (hold up). It is important to carefully follow the instructions on the top. Let me go over these: 0 = not observed; 1 = slight / ambiguous occurrence; 2 = definite occurrence, mild to moderate intensity, less than three minutes in duration; 3 = definite occurrence, severe intensity, greater than three minutes duration.

CBO Group

Now that we are finished watching the videos, you are ready to evaluate this child. Take out the form labeled "Rating form", this form asks you to estimate the percent engaged in each behavior and your confidence in that estimate. The percent engaged estimates appear on your computer screen. Please take a few minutes to fill out this form.

Next, I want you to fill out the "Global Rating Form" which asks your opinion on general aspects of this child's behavior. It looks like this (hold up) Please take a few minutes to fill out this form.

The final thing that you'll do today is to fill out a more detailed form asking specific questions about his child's behavior. The form looks like this (hold up). It is important to carefully follow the instructions on the top. Let me go over these: 0 = not observed; 1 = slight / ambiguous occurrence; 2 = definite occurrence, mild to moderate intensity, less than three minutes in duration; 3 = definite occurrence, severe intensity, greater than three minutes duration.

NAR Group

Now that we are finished watching the videos, you are ready to evaluate this child. Take out the form labeled "Rating form", this form asks you to estimate the percent engaged in each behavior and your confidence in that estimate. You may use your written data as a reference. Please take a few minutes to fill out this form.

Next, I want you to fill out the "Global Rating Form" which asks your opinion on general aspects of this child's behavior. It looks like this (hold up) Please take a few minutes to fill out this form.

The final thing that you'll do today is to fill out a more detailed form asking specific questions about his

child's behavior. The form looks like this (hold up). It is important to carefully follow the instructions on the top. Let me go over these: 0 = not observed; 1 = slight / ambiguous occurrence; 2 = definite occurrence, mild to moderate intensity, less than three minutes in duration; 3 = definite occurrence, severe intensity, greater than three minutes duration.

NM Group

Now that we are finished watching the videos, you are ready to evaluate this child. I want you to fill out the "Global Rating Form" which asks your opinion on general aspects of this child's behavior. It looks like this (hold up) Please take a few minutes to fill out this form.

The final thing that you'll do today is to fill out a more detailed form asking specific questions about his child's behavior. The form looks like this (hold up). It is important to carefully follow the instructions on the top. Let me go over these: 0 = not observed; 1 = slight / ambiguous occurrence; 2 = definite occurrence, mild to moderate intensity, less than three minutes in duration; 3 = definite occurrence, severe intensity, greater than three minutes duration.

Appendix O

Debriefing

Now that you are finished, I would like to debrief you on this study. But first, I want you to make sure that your participation code is on all of the sheets. Remember, this code is written on the outside of your packet and the last digit of the code is the number of the computer you sat at. Gather all of the materials you have filled out and place them in the packet and leave them on the desk, I will pick them up later.

Now I can tell you that this study investigated 2 independent variables, each with three levels, and their effect on the dependent variable, which was the accuracy of your observations. The first independent variable was the method by which you observed the girl in the video. Other groups were instructed to take notes in different ways than your group. The second independent variable was the information you received prior to the observation. Other groups were told different things about this child than your group. My hypothesis is that both the method and the introductory information will affect the accuracy of your observations and behavior

ratings. Your participation in this study will offer important information to psychologists about how to conduct a behavior observation. I thank you for participating in this study, I will post the results as soon as I can on your class web-site. You are free to leave, and again, I thank you for your time.

Bibliography

Achenbach, T.M., & Edelbrock, C. (1983). Manual for the Child Behavior Checklist and Revised Child Behavior Profile. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T.M., & Edelbrock, C. (1986). Manual for the Teacher's Report Form and teacher version of the Child Behavior Profile. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T.M. (1991). Manual for the Child Behavior Checklist/4-18 and 1991 profile. Burlington: University of Vermont, Department of Psychiatry.

Arkes, H.R., & Harkness, A.R. (1980). The effect of making a diagnosis on subsequent recognition of symptoms. Journal of Experimental Psychology: Human Learning and Memory, 6, 568-575.

Arkes, H.R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. Journal of Consulting and Clinical Psychology, 49, 323-330.

Arnoult, L.H. & Anderson, C.A. (1988). Identifying and reducing causal reasoning biases in clinical practice. In D.C. Turk & P. Salovey (Eds.), Reasoning, inference, and judgment in clinical psychology (pp.209-232). New York: Free Press.

Bakeman, R., & Gottman, J.M. (1986). Observing Interaction: An Introduction to Sequential Analysis. Cambridge, U.K.: Cambridge University Press.

Bargh, J.A. (1994) The Four Horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In R.S. Wyer, Jr., & T.K. Srull (Eds.), Handbook of Social Cognition (2nd Ed., pp.1-40). Hillsdale, NJ:Erlbaum.

Bargh, J.A., & Chartrand, T.L. (1999). The unbearable automaticity of being. American Psychologist, 54, 462-479.

Barnett, D.W., Lentz, F.E., & MacMann, G. (2000). Psychometric qualities of professional practice. In Shapiro, E.S., & Kratochwill, T.R. (Eds.) Behavioral Assessment in Schools: Theory, Research, and Foundations (2nd Ed.). (pp. 355-386). New York: Guilford.

Bergan, J.R., & Kratochwill, T.R. (1990). Behavioral Consultation and Therapy. New York: Plenum Press.

Bergan, J.R. (1995). Evolution of a Problem-Solving Model for Consultation, Journal of Educational and Psychological Consultation, 6, 111-123.

Bijou, S.W., Peterson, R.F., & Ault, M.J. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal of Applied Behavior Analysis, 1, 175-191.

Chapman, L.J., & Chapman, J.P. (1967). The genesis of popular but erroneous psychdiagnostic observations. Journal of Abnormal Psychology, 72, 193-204.

Chapman, L.J., & Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid diagnostic signs. Journal of Abnormal Psychology, 74, 271-280.

Ciminero, A.R., Calhoun, K.S., & Adams, H.E. (Eds.). (1986) Handbook of Behavioral Assessment New York: Wiley.

Cohen, J. (1960). A coefficient of interjudge agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cone, J.D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. Behavior Therapy, 9, 882-888.

Cone, J.D. (1981). Psychometric considerations. In Hersen, M., & Bellack, A.S. (Eds.) Behavioral Assessment: A Practical Handbook (2nd ed.) (pp.38-70). Pergamon: New York.

Cooper, J.O. (1987). Measuring and recording behavior. In J.O. Cooper, T.E. Heron, & W.L. Heward (Eds.), Applied Behavior Analysis (pp. 59-80). Columbus, OH: Merrill Publishing.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. Science, 243, 1668-1674.

Eiler, J.M., Nelson, W.W., Jensen, C.C., & Johnson, S.P. (1989). Automated data collection using bar code. Behavior Research Methods, Instruments, and Computers, 21(1), 53-58.

Einhorn, J.H., & Hogarth, R.M. (1978). Confidence in judgment: Persistence of the illusion of validity. Psychological Review, 85, 395-416.

Emerson, E., Reeves, D.J., & Felce, D. (2000). Palmtop computer technologies for behavioral observation research. In Thompson, T., Felce, D., & Symons, F.J. (Eds.) Behavior Observation: Technology and Applications in Developmental Disabilities (pp.47-60) Baltimore: Paul Brooks.

Ericcson, K.A., & Simon, H.A. (1980). Verbal reports as data. Psychological Review, 87, 215-251.

Faust, D., & Nurcombe, B., (1989). Improving the accuracy of clinical judgment. Psychiatry, 52, 197 - 208.

Faust, D., Hart, K., Guilmette, T.J., & Arkes, H.R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. Professional Psychology: research and Practice, 19, 508-515.

Fischhoff, B. (1982). Debiasing, In d. Kahneman, P.Slovic, & A.Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp.422-444). Cambridge: Cambridge University Press.

Forney, K.A., Leete, A.J. & Linburg, D.G. (1991). A bar code scoring system for behavioral research. American Journal of Primatology, 23, 127-135.

Foster, S.L., & Cone, J.D. (1995). Validity issues in clinical assessment. Psychological Assessment Special Issue: Methodological issues in psychological assessment research, 7(3), 248-260.

Gambrill, E.D. (1990). Critical thinking in Clinical Practice. San Francisco: Jossey-Bass.

Garb, H. N. (2000). Computers will become increasingly important for psychological assessment: Not that there's anything wrong with that! Psychological Assessment, 12, 31 - 39.

Garland, A.F. (1995). Pathways to adolescent mental health services: Adolescent help-seeking and teacher identification and referral. Dissertation Abstracts International Section B: The Sciences and Engineering. Vol 55 (7 - B), pp.3013.

Gilbert, D.T., & Osborne, R.E., (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. Journal of Personality and Social Psychology, 57, 940-949.

Glass, G.V., Willson, V.L., & Gottman, J.M. (1975). Design and Analysis of Time-Series Experiments. Boulder, CO: Colorado Associated University Press..

Gauron, E.F., & Dickinson, J.K. (1969). The influence of seeing the patient first on diagnostic decision-making in psychiatry. American Journal of Psychiatry, 126, 199-205.

Goodenough, F.L. (1949). Mental Testing, its History, Principles, and Applications. Rinehart: Oxford, England.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12, 19-30.

Guess, D., Roberts, S., Siegel-Causey, E., Ault, M., Guy, B., & Thompson, B. (1993). Analysis of behavior state conditions and associated environmental variables among students with profound handicaps. American Journal of Mental Retardation, 97, 634-653.

Hall, S., & Oliver, S. (1997). A graphical method to aid the sequential analysis of observational data. Behavior Research Methods, Instruments, and Computers, 29, 563-573.

Harris, M.J., (1994). Self-fulfilling prophecies in the clinical context: Review and implications for clinical practice. Applied and Preventive Psychology, 3, 145-158.

Hawkins, R.P. (1991). Is social validity what we are interested in? Argument for a functional approach. Journal of Applied Behavior Analysis, 24, 205-213.

Hintze, J.M., & Shapiro, E.S. (1995). Systematic observation of classroom behavior. In Thomas, A. & Grimes, J. (Eds.) Best Practices in School Psychology (3rd Ed.), (pp.651-660). Washington, D.C: National Association of School Psychologists.

Johnson, S.M., & Lobitz, G.K. (1974). Parental manipulation of child behavior in home observation. Journal of Applied Behavior Analysis, 7, 23-32.

Kahneman, D. (1973). Attention and Effort. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

Kahng, S., & Iwata, B. (1998). Computerized systems for collecting real-time observational data. Journal of Applied Behavior Analysis, 31, 253-261.

Kahng, S., & Iwata, B. (2000). Computer systems for collecting real-time observational data. In Thompson, T., Felce, D., & Symons, F.J. (Eds.) Behavior Observation: Technology and Applications in Developmental Disabilities (pp.35-46) Baltimore: Paul Brooks.

Kanfer, F.H., & Phillips, J.S. (1970). Learning foundations of Behavior Therapy. New York: Wiley.

Kazdin, A.E. (1998) Research Design in Clinical Psychology, 3rd Ed. New York: Allyn and Bacon.

Keppel, G. (1991). Design and Analysis: A Researcher's Handbook (3rd Ed.). New Jersey: Prentice Hall.

Lerner, J.S., & Tetlock, P.E. (1999). Accounting for the effects of accountability. Psychological Bulletin, 125, 255-275.

Mash, E.J., & Terdal, L.G. (Eds.). (1988). Behavioral Assessment of Childhood Disorders. New York: Guilford.

Meehl, P. (1973). Why I do not attend case conferences. In Psychodiagnosis: Selected Papers. (pp. 225-302). Minneapolis: University of Minnesota Press.

Meehl, P. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. American Psychologist, 50, 266-275.

Morgan, D.L., & Morgan, R.K. (2001). Single participant research design: Bringing science to managed care. American Psychologist, 56, 119-127.

Mischel (1968) Personality and Assessment. New York: Wiley.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Nelson, R.O., & Hayes, S.C. (1979). The nature of behavioral assessment: A commentary. Journal of Applied Behavior Analysis, 12, 491-500.

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and Shortcomings of Social Judgment. Englewood Cliffs: Prentice Hall.

Noldus, L.P., Trienes, R.J., Hendriksen, A.H., Jansen, H., & Jansen, R.G. (2000). The Observer Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. Behavior Research Methods, Instruments, & Computers, 32, 197-206.

O'Reilly, C., Northcraft, G., & Sabers, D. (1989). The confirmation bias in special education eligibility decisions. School Psychology Review, 18, 126-135.

Reed, M.L., & Edelbrock, C. (1983). Reliability and validity of the direct observational form of the Child Behavior Checklist. Journal of Abnormal Child Psychology, 11, 521-530.

Reschly, D.J., & Ysseldyke, J.E. (1995). School psychology paradigm shifts. In Thomas, A. & Grimes, J. (Eds.) Best Practices in School Psychology (3rd Ed.), (pp.17-31). Washington, D.C.: National Association of School Psychologists.

Sandman, C. A., Touchette, P.E., Ly, J., Marion, S. D., & Bruinsma, Y. E. (2000). Computer-assisted assessment of treatment effects among individuals with developmental disabilities. In Thompson, T., Felce, D., & Symons, F.J. (Eds.) Behavior Observation: Technology and Applications in Developmental Disabilities (pp.271-296) Baltimore: Paul Brooks.

Saunders, R.R., Saunders, M.D., & Saunders, J.L. (1994). Data collection with bar code technology. In T.I. Thompson & D.B. Gray (Eds.), Destructive behavior in developmental disabilities: Diagnosis and treatment (pp.102-116). Thousand Oaks, CA: Sage.

Shapiro, E.S. (1987). Behavioral assessment in school psychology. Hillsdale, NH: Erlbaum.

Shapiro, E.S. (1996) Academic skills problems: Direct assessment and intervention, (2nd Ed.). New York: Plenum.

Shapiro, E.S., & Kratochwill, T.R. (2000a) Introduction: Conducting a multidimensional behavior assessment. In Shapiro, E.S., & Kratochwill, T.R.(Eds.) Conducting School-Based Assessments of Child and Adolescent Behaviors (pp. 1-20). New York: Guilford.

Shapiro, E.S., & Kratochwill, T.R. (2000b) Conceptual foundations of behavioral assessment in schools. In Shapiro, E.S., & Kratochwill, T.R.(Eds.) Behavioral Assessment in Schools: Theory, Research, and Foundations (2nd Ed.). (pp. 3-18). New York: Guilford.

Skiba, R.J., Peterson, R.L., & Williams, T. (1997). Office referrals and suspension: disciplinary intervention in middle schools. Education and Treatment of Children, 20, 295-315.

Skinner, C.H., Rhymer, K.N., & McDaniel, E.C. (2000a). Naturalistic direct observation in educational settings. In Shapiro, E.S., & Kratochwill, T.R.(Eds.) Conducting School-Based Assessments of Child and Adolescent Behaviors (pp. 21-54). New York: Guilford.

Skinner, C.H., Dittmer, K.I., & Howell, L.A. (2000b). Direct observation in school settings: Theoretical issues. In Shapiro, E.S., & Kratochwill, T.R.(Eds.) Behavioral Assessment in Schools: Theory, Research, and Foundations (2nd Ed.). (pp. 19-45). New York: Guilford.

Snyder, M., Tanke, E.D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. Journal of Personality and Social Psychology, 35, 656-666.

Snyder, M., & Swann, W.B., Jr. (1978). Hypothesis-testing processes in social interaction. Journal of Personality and Social Psychology, 35, 656-666.

Snyder, M., & Thomsen, (1988). Interactions between therapists and clients: Hypothesis testing and behavioral confirmation. In D.C. Turk & P. Slovey (Eds.), Reasoning, Inference, and Judgment in Clinical Psychology. New York: Free Press.

Steege, M.W., & Wacker, D.P. (1995). Evaluating the effectiveness of applied interventions. In Thomas, A. & Grimes, J. (Eds.) Best Practices in School Psychology (3rd Ed.). (pp.625-636). Washington, D.C.: National Association of School Psychologists.

Strosahl, K.D., & Linehan, M.M. (1986). Basic Issues in Behavioral Assessment. In Ciminero, A.R., Calhoun, K.S., & Adams, H.E. (Eds.) Handbook of Behavioral Assessment (2nd Ed.) New York: Wiley.

Suen, H.K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification, Behavioral Assessment, 10, 343-366.

Suen, H., & Lee, P. (1985). Effects of the use of percentage agreement on behavioral observation reliabilities: A reassessment, Journal of Psychopathology and Behavioral Assessment, 7, 221-234.

Temerlin, M.K., & Trousdale, W.W. (1969). The social psychology of clinical diagnosis. Psychotherapy: Theory, Research, and Practice, 6, 24-29.

Thompson, T., Felce, D., & Symons, F.J. (2000). Principles of Behavior Observation: Assumptions and strategies. In Thompson, T., Felce, D., & Symons, F.J. (Eds.) Behavior Observation: Technology and Applications in Developmental Disabilities (pp.3-16) Baltimore: Paul Brooks.

Tufte, E. (1990). The Graphical Display of Quantitative Information. Cheshire, CT: Graphics Press.