EXAMINING THE INFLUENCE OF INTERVAL AND OBSERVATION LENGTH ON THE DEPENDABILITY OF DATA

A dissertation presented by

Tyler David Ferguson

Submitted to
The Department of Counseling and Applied Educational Psychology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the field of

School Psychology

Northeastern University
Boston, Massachusetts
January, 2014
TABLE OF CONTENTS

Title Page ................................................................................................................................. 1
Table of Contents ...................................................................................................................... 2
List of Tables ............................................................................................................................. 4
List of Figures ........................................................................................................................... 5
Acknowledgements .................................................................................................................. 6

CHAPTER 1: LITERATURE REVIEW

Abstract .................................................................................................................................. 7

Direct Observation in Schools: A Review of the Psychometric Literature ....................... 8
  Methods of Observation .......................................................................................................... 9
    Naturalistic observation ......................................................................................................... 9
    Systematic direct observation ............................................................................................ 10
    Time-sampling methods ...................................................................................................... 12

Psychometric Properties of Assessment Data ...................................................................... 13
  Classical Test Theory .......................................................................................................... 15
    Problems with classical test theory .................................................................................... 16
  Generalizability Theory ....................................................................................................... 16
  Psychometric Properties of Direct Observation Data ......................................................... 17

Factors Influencing the Psychometric Adequacy of SDO Data ............................................. 20
  Behavioral Definition .......................................................................................................... 22
  Scheduling of Observation .................................................................................................... 23
  Time Sampling Methods ....................................................................................................... 26
  Interval Length ..................................................................................................................... 29
  Observation Length ............................................................................................................. 31
  Number of Observations ...................................................................................................... 34

Conclusions and Implications for Research and Practice ....................................................... 39

REFERENCES ......................................................................................................................... 42

CHAPTER II: EMPIRICAL RESEARCH ARTICLE

Abstract .................................................................................................................................. 48

Introduction ............................................................................................................................... 49
  Interval length ........................................................................................................................ 52
  Observation length ................................................................................................................ 54
  Purpose of Study ................................................................................................................... 55

Method ................................................................................................................................... 56
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>56</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>57</td>
</tr>
<tr>
<td>Procedure</td>
<td>58</td>
</tr>
<tr>
<td>Systematic direct observation</td>
<td>58</td>
</tr>
<tr>
<td>Observer training</td>
<td>59</td>
</tr>
<tr>
<td>Interobserver agreement</td>
<td>59</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>59</td>
</tr>
<tr>
<td>Results</td>
<td>61</td>
</tr>
<tr>
<td>Discussion</td>
<td>64</td>
</tr>
<tr>
<td>Limitations</td>
<td>68</td>
</tr>
<tr>
<td>Implications for Research and Practice</td>
<td>69</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>72</td>
</tr>
<tr>
<td>TABLES</td>
<td>76</td>
</tr>
<tr>
<td>FIGURES</td>
<td>80</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Mean (SD) Student Engagement using 5-sec Time Sampling by 5-min Block .......... 76
Table 1 (cont.). Mean (SD) Student Engagement using 5-sec Time Sampling by 5-min Block.. 77
Table 2. G Study Results for the Full Model (p x o) by Observation and Interval Length ........ 78
Table 3. D Study Results for the Full Model (p x o) by Observation and Interval Length ........ 79
LIST OF FIGURES

Figure 1. Percentage of variance attributable to person by observation and interval length........ 80
ACKNOWLEDGEMENTS

There are a number of people that have supported me during my time at Northeastern that I would like to thank. First, I would like to thank my classmates as well as the staff and faculty from the Department of Counseling and Applied Educational Psychology. Coming to campus often felt like spending time with friends as much as it did working towards my degree requirements, and I have many memories of our good times together. Next, I would like to thank my friends and family. They have supported me throughout all of my endeavors, and I know I never could have gotten this far without them. Most of all, I would like to thank my adviser and mentor, Amy Briesch. I have learned so much from you during our time working together, and you have always made me feel more like a colleague than an advisee. You have always had my best interest in mind, and I am grateful for your patience and guidance these past five years. Although I am sure that we will continue to collaborate professionally, more than anything, I hope that we can move from being colleagues to being friends. Thank you, everyone, for all of your support.
CHAPTER 1

Abstract
Direct observation techniques are popular among school psychologists, and a number of methods can be employed. Despite its popularity, the psychometric properties of systematic direct observation have not received much attention. Because procedures and targets vary widely across coding systems, research in this area provides insight into how procedural aspects influence the reliability of direct observation data. Specifically, the influences of behavioral definition, time sampling method, interval length, observation length, time of observation, and number of observations on the obtained results of systematic direct observation are explored. Both implications for researchers and practitioners are discussed.
LITERATURE REVIEW

Direct Observation in Schools: A Review of the Psychometric Literature

Since the inception of school-based behavioral assessment methods, direct observation has been a foundational technique (Landau & Swerdlik, 2005). Direct observation methods usually involve an individual other than the teacher entering a setting of interest to observe and record the behavior of a particular student (Merrell, 2008). Data collected in this fashion are less influenced by bias than information obtained from teachers or other professionals, which may prove overly subjective (Shapiro, 2004). Indeed, many researchers argue that direct observation methods are superior to other assessment methods because they are more objective, less biased, and can be conducted in the individual’s natural setting (Merrell, 2008). In addition, information obtained through direct observation is broadly useful, in that it can be used to inform placement decisions as well as intervention planning and monitoring.

Research has shown that direct observation is one of the most popular behavioral assessment methods used by school psychologists to assess student behavior. In a survey of 1,000 school psychologists, Wilson and Reschly (1996) found that structured observation methods were the most frequently used assessment tool. Practitioners reported conducting over 15 behavioral observations of student behavior over the course of a typical month. Direct observation has also been identified by Shapiro and Heick (2004) as an assessment method used frequently by school psychologists to investigate social/emotional/behavioral referral concerns. Research by Stinnett, Harvey, and Oehler-Stinnett (1994) demonstrated similar findings, with NASP members reporting that they conduct approximately 10 observations in a given month, and that structured observation methods (i.e., event, interval, and time-sampling) were among the top ten assessment methods used.
One reason direct observation may be popular is its utility in assessing problem behaviors. Direct observation has been implicated as a vital component in the assessment of emotional and behavioral disorders as well as functional behavioral assessment (McConaughy & Ritter, 2008; Steege & Watson, 2008). Hintze, Volpe, and Shapiro (2002) described four advantages of direct observation, including: (a) informing the importance or social validity of a behavior by noting its frequency in the natural environment, (b) systematically examining relationships between the behavior and its antecedents and consequences, (c) informing the development of testable hypotheses regarding the function(s) of the behavior, and (d) providing insight into the function of a behavior rather than focusing on topographical and descriptive accounts of what occurred. Through the use of direct observation, valuable information can be collected to inform the assessment and intervention processes in an objective way.

**Methods of Observation**

In general, there are two approaches practitioners can take to collect direct observation data: naturalistic observation or systematic direct observation. Compared to systematic direct observation techniques, naturalistic observation is used almost twice as often by school psychology practitioners (Wilson & Reschly, 1996). This may be related to ease in use and minimal training requirements when compared to learning a systematic observation coding system (Hintze et al., 2002).

**Naturalistic observation.** Naturalistic observation techniques, sometimes referred to as narrative or anecdotal observation, involve observing a person in his or her natural environment and recording behaviors of interest as they occur (Hintze et al., 2002). This method provides a general impression of what took place during the observation session. Generally, behaviors are recorded chronologically as they occur in the natural environment. A major problem with
naturalistic observation techniques, however, is the limited utility of the data, which are
 descriptive in nature and only allow for summary statements to be made about what was
 observed (Hintze et al., 2002). Observers may also pay more attention to behaviors that align
 with their preconceived notions of an individual, drawing attention away from behaviors that
 may provide conflicting evidence. These confirmatory search strategies may tarnish the value of
 observational data by giving an observer an inaccurate impression of what took place (Gilovich,
 1993). Despite these concerns, data gathered from anecdotal observation methods can be used to
develop testable hypotheses about a student's behavior, which can be further investigated
through other methods.

**Systematic direct observation.** In contrast to naturalistic observation, systematic direct
observation techniques are used to quantify specific behaviors of interest using predetermined
procedures (Hintze et al., 2002). Salvia and Ysseldyke (2004) described five characteristics of
systematic direct observation: (a) the goal is to measure specific target behaviors; (b) the target
behaviors are operationally defined in a clear manner before observations occur; (c) observations
take place using standardized procedures and are conducted in an objective manner; (d) times
and locations for observation are purposefully selected and specified; and (e) standardized
procedures are used to score and summarize data across observers. Because data are collected in
a defined manner and are quantitative in nature, they can be directly compared across
observations and are therefore appropriate to inform decision-making processes.

A vital component of systematic direct observation procedures is clearly defining the
target variables. An operational definition should allow the observer to accurately determine
whether or not a behavior occurred. Hawkins and Dobes (1977) offer guidelines for developing
explicit behavioral definitions. First, the definition should only refer to observable characteristics
of the behavior and environment, remaining as objective as possible. Second, the definition should be readable and unambiguous, allowing any observer to accurately understand and paraphrase it. Finally, the definition should be complete, providing clear criteria of what is and is not the behavior. These aspects are important in helping the observer discriminate the target behavior from other similar responses.

Once the behavior has been operationally defined, characteristics of interest are used to determine the methods of data recording (Hintze, 2005). Observational procedures can be applied to specific behavioral events or states of behavior, with each type of behavior necessitating different methods of data collection (Merrell, 2008). Event recording, also referred to as frequency recording, is used to measure how many times a specific behavior occurs over the course of an observation session. Merrell (2008) described three important criteria to consider when using event recording methods: (a) the behavior must have a clear beginning and end; (b) the duration of each behavioral event should be approximately the same; and (c) the behavior should not occur so frequently that it is difficult to separate each occurrence. Some examples of behaviors that may be appropriate for frequency recording methods are talking out of turn and hand-raising.

State behaviors, which represent conditions that are more prolonged than event behaviors, can be investigated using a few different data recording methods, including duration, latency, and time-sampling recording (Merrell, 2008). Duration recording, as its name implies, involves recording the length of time a behavior lasts for. Duration recording can be particularly helpful in situations where changing the duration of a behavior is an important target for intervention, such as temper tantrums or studying. Latency recording should be used when the length of time between the opportunity to perform a behavior and how long it takes to actually
begin is of interest, such as recording the delay between when a teacher asks a student to take out his or her homework and when he or she actually does so.

**Time-sampling methods.** Although there are a number of methods that can be used to quantify behavior, one of the most popular observational targets is engagement due to the important implications it has in the learning process (Greenwood, 1996). When interested in describing behaviors such as engagement that occur moderately to steadily, time-sampling methods are most appropriate. Time sampling involves dividing the observation period into a number of equal intervals and recording whether or not a specific target behavior occurs within each interval (Hintze et al., 2002). This approach differs from frequency, duration, and latency recording in that it only approximates the amount of time a behavior occurs rather than providing an exact measurement of an observed behavior. A number of different coding procedures can be used, and it is possible to record behaviors under different schemes simultaneously. There are three distinct time sampling methods: (a) whole-interval recording, which scores a behavior as being present if it occurs for the entire duration of an interval; (b) partial interval coding, which scores a behavior as being present if it occurs at any time during the interval; and (c) momentary time sampling, which records the occurrence or nonoccurrence of a behavior only at the moment the time interval begins or ends.

Because different criteria are used by each time-sampling method to code target behaviors, measurement error associated with each method also differs. For intervals in which the target behavior does not occur, all three time-sampling methods, barring human error, will be scored the same (i.e., as a nonoccurrence). The frequency and duration of a behavior play an important role in how the occurrence of a target is coded when using these three methods (Ary, 1984). Because partial interval methods code the occurrence of a behavior any time it takes place
during an interval, the number of intervals coded as an occurrence increases as the frequency of
the target behavior increases. Therefore, sessions in which behaviors are frequently initiated and
ended will produce greater proportions of occurrences than a session with the same actual
duration of behavior but with fewer changes. This may lead to an overestimation of the true rates
of a behavior (Ary, 1984). In contrast, whole-interval recording methods are negatively impacted
by frequency because the more often a behavior is initiated and ended, the more intervals there
are in which a behavior was not sustained long enough to be coded as an occurrence. Duration
plays a much more important role for whole-interval methods, as the behavior needs to last the
entire interval to be coded as an occurrence. These factors may cause whole-interval recording to
underestimate true rates of behavior (Ary, 1984). The results of momentary time-sampling
methods are not systematically related to frequency, but random error will increase as the
number of intervals containing behavior change increases (Ary, 1984). The outcome of all three
time-sampling methods will be similar when behavior changes occur infrequently. However, as
frequency increases, the outcomes of each method become increasingly divergent from each
other. Due to these differences, it is important to consider the frequency and duration of a target
behavior when selecting a particular time-sampling approach.

**Psychometric Properties of Assessment Data**

Regardless of the method being used to collect data, most assessment techniques allow
practitioners and researchers to obtain numerical descriptions of the degree to which an
individual demonstrates a particular characteristic or behavior (Nunnally & Bernstein, 1994).
These data are only useful for descriptive or analytic purposes, however, if they are reliable and
valid. The American Psychological Association (APA) and National Association of School
Psychologists (NASP) emphasize these principles, urging practitioners to use evidence-based
methods of assessment. Federal law is also clear on the issue: the Individuals with Disabilities Education Improvement Act (IDEIA, 2004) mandates that assessment procedures be validated for their intended purposes and possess adequate psychometric properties. Therefore, practitioners are obligated to use assessment tools that are psychometrically sound.

Whether or not the obtained numbers accurately represent the behavior of interest is conceptualized as the validity of the data. Validity can be explored in a number of ways, including content validity (i.e., the degree to which items encompass the entire range of responses for a given construct), criterion-related validity (i.e., how similar results are to those of other instruments used to measure the same construct), and construct validity (i.e., the extent to which the instrument measures the intended construct) (Silva, 1993). Messick (1995) discussed some common threats to validity, including construct underrepresentation and construct irrelevant variance. Construct underrepresentation occurs when the assessment is too narrow and neglects to incorporate important dimensions or aspects of the construct. For example, a measure of reading fluency that assesses rate but not accuracy would neglect an important part of this construct (Fuchs, Fuchs, Hosp, & Jenkins, 2001). On the other hand, construct irrelevant variance occurs when the assessment is too broad in its scope. Using the example above, incorporating reading comprehension into a measure of reading fluency also would not properly represent this construct as comprehension and fluency are separate entities.

For an instrument to be considered valid, it must also demonstrate reliability. Reliability concerns the consistency of a measure in assessing a particular construct (Crocker & Algina, 2006). Reliability can be assessed in a number of ways depending on what type of consistency is being examined. When interested in examining the degree to which items consistently assess a given construct, either alternate form reliability (i.e., examining the similarity between two
different forms of a test) or internal consistency (i.e., how similar items within a test are consistent with one another) can be assessed (Crocker & Algina, 2006). To determine how consistent data are over time, test-rest reliability can be calculated, which compares scores on the same test given at different points in time (Crocker & Algina, 2006). To assess consistency based on who is collecting the data, interobserver reliability describes the consistency of data between different observers and intraobserver reliability describes how consistently a single observer rates the same items (Crocker and Algina, 2006). There are a number of different factors that may affect the reliability of data. The number of test items used has been demonstrated to have an impact on reliability, with reliability increasing as the number of items also increases because there is more evidence available to approximate a value (Thorndike, 2005). The amount of time between measurements has also been theorized to impact reliability, with measurements taken across a small period of time yielding more consistent results than those obtained over longer durations (Salvia & Ysseldyke, 2007). Training effects may also play an important role in influencing the reliability of data, with observers who are trained or experienced with a particular data collection tool yielding more reliable results compared to those individuals with less training or experience (Salvia & Ysseldyke, 2007).

**Classical Test Theory**

Historically, technical adequacy has been investigated through the lens of classical test theory (CTT). Under this model, each piece of data, known as an observed score, is partitioned into two components: true score and error score (Cone, 1998; Suen & Ary, 1989). Classic concepts of reliability are based on the idea that what is being measured represents a sample of a “true” score. Data are considered reliable to the extent that observed scores are similar to true scores. In other words, the more reliable a score is, the more closely it represents the actual
manifestation of a given trait or behavior (Cone, 1977; Salvia & Ysseldyke, 2004). Because the values of true scores and error scores are not known, CTT ignores true scores when determining reliability by estimating their relative proportion through the use of observed scores. This is accomplished by obtaining two scores for a given target that meet parallel test assumptions, meaning that both measurements are designed to assess the same “true” score (i.e., test-retest, split-half reliability). Because parallel tests are considered to have identical true scores, observed scores can be compared to determine the stability of scores.

**Problems with classical test theory.** Although CTT is a popular method of investigating technical adequacy, there are several limitations to this approach. Because CTT breaks variance down into only two sources, true score and error, only one source of error can be estimated at a time (Shavelson & Webb, 1991). Error is known to be influenced by a number of factors, but because error is not differentiated, this information is lost when using a CTT approach (Shavelson, Webb, & Rowley, 1989). Shavelson and Webb (1991) also cautioned that CTT could not be used when one is interested in investigating the reliability of data for the purposes of absolute decision (e.g., intra-individual decisions). This is because CTT assumes that a true score is a stable construct, which may be misleading when assessing a construct that may be variable, such as behavior (Silva, 1993). For example, if data are collected to monitor treatment progress, improvements or deteriorations in behavior may be mistakenly attributed to error.

**Generalizability Theory**

An alternative approach that can be used to assess the psychometric properties of data is Generalizability (G) theory. Based on domain sampling, G theory postulates that any single observation of a behavior represents a random sampling of that behavior from a hypothetical domain of all possible contexts. G theory is used to assess the extent to which a set of
measurements generalize to the actual values of a given target (Cone, 1977). The dependability of a measure, which incorporates both reliability (i.e., consistency of target under specified conditions) and validity (i.e., extent to which the observation reflects an actual target over all possible conditions), is also investigated through G theory.

There are a few notable advantages to using G theory to assess the reliability of data over CTT approaches. Unlike in CTT, where all error variance is pooled together, G theory allows one to estimate the proportion of error associated with a particular facet of a given measurement. For example, in the case of direct observation, the proportion of error attributable to duration, observer, and setting can be calculated simultaneously (Cone, 1977). This allows for close examination of the relative magnitude of each source of error, allowing for those contributing the largest portions to be identified. This knowledge could be used to examine existing psychometric properties as well as to improve future assessment procedures. For example, the error associated with the facet of rater could be examined to see how variability in scores is explained by using different observers. If a large portion of error is associated with rater, it may be important to make sure observers are coding data in the same way, which may involve additional training.

Studies can also be used to derive generalizability and dependability coefficients to help inform decision-making processes. To examine within-individual decisions, where the performance of others is not taken into account (i.e., formative assessments), dependability coefficients are used to determine reliability. In contrast, generalizability coefficients are used to evaluate reliability when between-individual decisions are of interest, such as in the case of norm referenced assessments (Shavelson & Webb, 1991).

**Psychometric Properties of Direct Observation Data**
Although we know that establishing the psychometric properties of assessment tools is a critical task, there have been conflicting views over whether or not psychometric principles can be applied to systematic direct observation (Silva, 1993). These differences of opinion are rooted in the debate over whether or not psychometric considerations apply to behavioral assessment methods in general. Because systematic direct observation focuses on behaviors that are considered variable and situation specific, some authors argue that by their very nature direct observation data are not consistent (Nelson, Hay, & Hay, 1977). Inconsistency or instability of data may actually reflect inconsistent or unstable behavior, which may be clinically relevant to consider and explain rather than simply viewing as measurement error (Silva, 1993). However, it has also been argued that the technical adequacy of systematic direct observation data should be scrutinized if used to inform decision-making (Cone, 1988), and federal law is in accordance with this mindset (IDIEA, 2004).

Hintze (2005) discussed seven psychometric considerations he believed were important to apply to systematic direct observation. First, internal consistency reliability can be used to examine whether or not an observation session is of sufficient length to adequately describe a behavior. Second, the concept of test-retest reliability can be applied to determine whether or not a sufficient number of observations have been conducted to examine the consistency of behavior over time. Third, interobserver agreement can be used to describe the extent to which two observers agree on the occurrence or nonoccurrence of a behavior. Content, concurrent/convergent, and predictive validity consider how well the data accurately describe the target behavior, agree with other sources of assessment information (e.g., interviews, behavior ratings), and accurately predict behavior in future situations, respectively. A final consideration, sensitivity to change, examines the extent to which behavioral observation data change as a
function of environmental manipulations and developmental changes over time. The extent to which each of these characteristics has been explored varies considerably in the systematic direct observation literature.

Of the psychometric considerations described by Hintze (2005), the most popular approach to assessing the technical adequacy of systematic direct observation data within the literature has been through interobserver agreement. As with other CTT methods, data collected by each observer are considered to represent parallel tests, which are compared to describe the consistency of the data across observers (Suen & Ary, 1989). Interobserver agreement may be estimated using a number of different metrics, including a smaller/larger index, percentage agreement index, occurrence and nonoccurrence agreement indices, coefficient kappa, and coefficient phi (Hintze, 2005). A number of direct observation coding systems have used interobserver agreement to validate psychometric adequacy, including the Academic Engaged Time Code (AET-SSBD), Attention-Deficit Hyperactivity Disorder School Observation Code (SOC), Behavioral Observation of Students in Schools (BOSS), Class Observation Code (COC), Direct Observation Form (DOF), State-Event Classroom Observation System (SECOS), and the Student Observation System (SOS) (Volpe, DiPerna, Hintze, & Shapiro, 2005). In addition, interobserver agreement is a common metric in observational research, and is used regularly as the primary indicator of reliability of the data collected through systematic direct observation methods.

Although it is a popular metric, there have been concerns raised over the appropriateness of using interobserver agreement solely to assess the reliability of systematic direct observation data. The fact that two observers agree on the number of occurrences or non-occurrences of a behavior over a specific time period does not provide any information as to whether or not their
data accurately portray what occurred (Johnston & Pennypacker, 1993). Furthermore, because there are only two possible outcomes when comparing data (i.e., agreement or non-agreement), it is possible that agreement occurs by chance. Hartmann (1982) likened this to two individuals repeatedly flipping a coin, which is bound to result in chance agreement over time. Another issue regarding the use of interobserver agreement to describe reliability is the fact that awareness of impending reliability checks has been shown to improve accuracy and reliability, further complicating the issue (Reid, 1970; Romanczyk, Kent, Diament, & O'Leary, 1973). Therefore, when interested in the investigating the technical adequacy of systematic direct observation data, interobserver agreement alone does not appear to be sufficient.

**Factors Influencing the Psychometric Adequacy of Systematic Direct Observation Data**

Regardless of the technique being used to investigate psychometric adequacy, the reliability and validity of an assessment instrument are considered to be tied directly to the testing situation due to the number of environmental, situational, and procedural factors that can influence data collection (Herbert & Attridge, 1975). When standardized procedures and instruments are used, as is the case with many measures of cognitive and academic achievement, reliability and validity are expected to be stable across administrations. This stability lends itself well to determining the technical adequacy of these assessment tools. In contrast, although the flexibility inherent to systematic direct observation is one of the main advantages of this assessment tool (Briesch & Volpe, 2007), this poses a particular problem for describing the psychometric properties of systematic direct observation, where assessment targets and methods can vary widely. For example, Karweit and Slavin (1981) found that several studies investigating the relationship between time-on-task behavior and academic achievement produced inconsistent results, which they attributed to variations in operational definitions and observation procedures.
used to collect time-on-task data. Their findings suggest that even when theoretically observing the same construct, the reliability and validity of systematic direct observation data are specific to the operational definition, particular coding system, and particular set of environmental conditions during the observation session. Although it would be theoretically possible to investigate the psychometric properties of each systematic direct observation procedure, this approach is simply not feasible. Therefore, it may be more helpful and practical to consider research on the psychometric properties of systematic direct observation data in the context of providing insight into how different procedural aspects affect reliability and validity to inform general guidelines for implementing systematic direct observation procedures.

A systematic literature review was conducted during the fall of 2012 to identify those studies that have investigated the psychometric properties of direct observation data and, more specifically, how the reliability/validity of observational data are influenced by procedural factors. Articles were obtained through literature searches using the ERIC, PsycInfo, and Dissertation Abstracts databases. Keywords included a combination of the following terms: direct observation, psychometrics, generalizability, reliability, and validity. Although many studies were identified in which reliability and/or validity were investigated for a specific commercially available coding system, these studies were not reviewed because the goal of this search was to identify those studies in which some procedural aspect of systematic direct observation was manipulated to inform how this factor influenced the obtained data. Ancestral searches were conducted within the articles that were identified during the initial database search. This review yielded 14 studies evaluating factors related to the psychometric properties of systematic direct observation data. Six major areas found to influence technical adequacy
were identified: behavioral definition, time sampling method, interval length, observation length, time of observation, and number of observations.

**Behavioral Definition**

A key element of any systematic direct observation code is developing an operational definition that can discriminate between the occurrence and nonoccurrence of an observational target. Despite being a critical component of a systematic direct observation code, only one study has specifically examined the influence of behavioral definition on the reliability of data.

Karweit and Slavin (1982) manipulated five aspects related to direct observational procedures, including operational definition, to examine their influence on the reliability of time on-task data. Students from 18 classes, grades 2 through 5, were selected based on results of the Comprehensive Test of Basic Skills (CTBS). Two students (e.g., one boy and one girl) were chosen from the top third, middle third, and lower third of each class based on CTBS results, for a total of 108 students. All data were collected during math class, which averaged 50 minutes in duration. Momentary time sampling procedures utilizing 30-second intervals were used to code three pieces of information: (a) the nature of the task (i.e., procedural, seatwork, lecture); (b) the student response to the task (i.e., on-task, off-task, or no task opportunity); and (c) the content of instruction (i.e., two-digit multiplication or homework review). Students were observed in each class using a predetermined order, and student responses were only coded during instructional parts of the class. “Other response” was coded if a student was not on-task but not off-task either, such as when sharpening a pencil or obtaining classroom materials. Two definitions of on-task were used: both were defined as behavior appropriate to the task at hand, but one was modified to permit momentary demonstrations of off-task behavior (i.e., gazing out the window or fidgeting). Small changes in rates of on-task behavior were found based on these different
definitions (i.e., 83.7% to 86.6% for grades 2/3, 81.5% to 84.5% for grades 4/5), although reliability data were not collected. The authors suggested that the results of systematic direct observations may be susceptible to seemingly arbitrary definitional decisions. When specifically applied to time-on-task, the authors suggested that the inclusion of momentary inattention in definitions of on-task behavior should be guided by views on learning processes.

This study demonstrates that even small nuances in operational definitions may have an impact on data outcomes. Researchers must carefully consider their theoretical perspective on a given construct when developing an operational definition. Also, although many studies may appear to investigate the same target, it may not be appropriate to make comparisons across these studies if the way a behavior is defined is not consistent. Convergent or divergent data may be better explained by differences in behavioral definition than true differences in study outcomes. For school-based practice, these findings suggest that operational definitions should be carefully constructed and used consistently across observational approaches.

**Scheduling of Observation**

The time and day an observation takes place may have an impact on data outcomes because behavior may be influenced by temporal factors. One clear example is within school systems, where the beginning and end of the school year as well as days immediately before or after holidays are known to produce behavioral change among students (Karweit & Slavin, 1982). There are three studies that have either directly or indirectly looked at the influence of scheduling on the reliability of data.

Among the other factors discussed that influence time-on-task data, Karweit and Slavin (1982) also investigated the phenomenon of scheduling of observations. They compared data collected from 10-day periods in February and May and found that mean scores and standard
deviations for time-on-task for each classroom were not very different across these two time periods, suggesting that observational period may not be very consequential. However, when comparing rates of time-on-task within a given day during the first 10 minutes, the second 10 minutes, and the third 10 minutes of instruction, data varied considerably for some classrooms across these time periods. This study suggests that although the general time period of the observation may not be consequential, the timing of an observation session within a given day may impact data outcomes.

Hintze and Matthews (2004) used G theory to investigate the dependability of systematic direction observation data across time and setting for levels of academic engagement. Participants were 14 students from an inclusionary fifth grade classroom. Students were observed for on-task/off-task behavior using a 15-second momentary time-sampling procedure over a 15 minute period. Data were collected twice a day (i.e. in the morning during math and afternoon during language arts) for ten consecutive school days. The greatest proportion of variance in scores was attributable to individual differences among persons (62%), followed by the error term (24%) and the interaction between person and setting (13%), suggesting that students were not consistent in their behavior across the morning and afternoon observation sessions. Because both the time of the observation as well as the classroom activity are incorporated into the facet of setting, however, it is unclear what role timing (i.e., morning versus afternoon) and activity (i.e., math versus language arts) uniquely played in influencing the dependability of data.

Although not their primary purpose, Volpe, McConaughy, & Hintze (2009) also explored the impact of scheduling on the reliability of systematic direct observation data through G theory. Twenty four students, ages 6- to 11-years old, who were referred by their teachers due to
learning and behavioral problems were observed across 18 different elementary schools. Participants were observed on eight occasions in the morning and afternoon across four different days using the Direct Observation Form (DOF; McConaughy & Achenbach, 2009). The DOF contains 89 items that describe behavior, which are rated using a 4-point Likert-type scale (i.e., 0 = no occurrence, 3 = definite occurrence with severe intensity, high frequency, or 3 or more minutes total duration) immediately following a 10-minute observation. Ratings on these items are used to derive five syndrome scales (i.e., Sluggish Cognitive Tempo, Immature/Withdrawn, Attention Problems, Intrusive, and Oppositional), a DSM-oriented ADHD scale with Inattention and Hyperactivity-Impulsivity subscales, and a Total Problems score based on all 89 items. The DOF also uses time-sampling methods to collect data about the on-task behaviors of a target during the last 5-seconds of each minute of observation, with on-task coded if they exhibit the behavior for a majority of the 5-second interval. Across all behavioral targets, the largest proportion of variance was attributable to the error term (Range = 51% - 87%), followed by Person (Range = 6% - 28%). The facets of Time of Day (Range = 0% - 3%), Time of Day nested within Occasion (Range = 0% - 13%), and the interaction between Person and Time of Day (Range = 0 – 18%) varied in their contribution across scales, indicating that the time of day in which data were collected influenced data outcomes but to different degrees depending on what was being measured. Unlike Hintze and Matthews (2004), where instructional arrangement was systematically altered across morning and afternoon observations, Volpe and colleagues (2009) did not control for instructional arrangements, which may limit the interpretation of these findings.

Although the results of these studies suggest that timing of observation may play a role in influencing the reliability of direct observation data, it is difficult to draw general conclusions
from this research. Because classroom activity was not systematically controlled in the latter studies, this factor is difficult to separate from the actual scheduling of the observation. However, in terms of providing insight into school-based concerns, this research provides some evidence that time-on-task behavior may be generally consistent over some periods of time, but classroom activity may play an important role in the reliability of data. More research is necessary in this area to explore this factor.

**Time Sampling Methods**

As discussed previously, choice of time-sampling method (i.e., partial-interval, whole-interval, and momentary time sampling) may influence the reliability of observational data because different criteria are used to code behaviors. Over the past four decades, several studies have been conducted to specifically examine the influence of time sampling method on the reliability of observational data.

This issue was first explored by Powell, Martindale, and Kulp (1975). Because whole-interval, partial-interval, and momentary time-sampling methods use different criteria for recording behavior, the authors were interested in determining whether or not they yield different results. The in-seat behavior of a secretary (i.e., posterior is in contact with the seat) was continuously coded and also coded using partial interval, whole interval, and momentary time sampling methods for six 20-minute periods using video footage. Momentary time sampling data were collected every 10, 20, 40, 80, 120, 240, 400 and 600 seconds, whereas partial- and whole-interval data were collected using intervals of 10, 20, 40, 80, and 120 seconds. The authors compared the results of time-sampled measures of each behavior to those obtained through continuous measurements. When time-sampling methods were conducted up through 120-second intervals, the momentary time-sampling and continuous measures agreed closely. If partial- or
whole-interval recording methods were used, discrepancies from the continuous measure were a function of interval length, with intervals up through 40 seconds for whole-interval and 80 seconds for partial-interval yielding acceptable levels of correspondence to the continuous measure. In general, whole-interval time-sampling methods were found to underestimate in-seat behavior and partial-interval methods were found to overestimate rates of in-seat behavior, whereas momentary time-sampling provided the closest approximation of behavior.

In a follow-up study, Powell, Martindale, Kulp, Martindale, and Bauman (1977) again investigated the measurement error associated with different time-sampling methods. Video footage of in-seat behavior was targeted through direct observation, but the in-seat behavior of the participant was engineered; that is, he was cued to be in his seat for 20%, 50%, or 80% of the session by a timer. This was done to determine how variability in behavior may influence outcomes. The results of time-sampled measures of each behavior were again compared to those obtained through continuous measurements. When behavior occurred at less steady rates (i.e., 20%), the error in time sampling methods was found to be large compared to when behavior occurred more frequently (i.e., 50-80%). Up through 60-second intervals, momentary time-sampling methods demonstrated close agreement with the continuous measure, regardless of behavior rate. In general, findings were similar to those of their previous study, which found that partial- and whole-interval time sampling methods tended to overestimate and underestimate rates of in-seat behavior, respectively, and momentary-time sampling methods produced less systematic distortion when measuring in-seat behavior.

Murphy and Goodall (1980) also examined how choice of time-sampling method affects the obtained data. Three brief duration (i.e., fist to chin, pinching self, and vocalization), medium duration (i.e., hand flapping, body and hand tapping/posturing), and long duration (i.e., body
rocking, finger sucking, and hand posturing) behaviors of eight intellectually impaired children were coded using four different time sampling methods (i.e., 2.5-second partial-interval, 10-second partial interval, 10-second whole-interval, 10-second momentary time-sampling methods), and the duration of each behavior was recorded using a Rustrak event recorder. The extent to which each recording method produced error was calculated by subtracting the estimates of behavioral duration provided through each time sampling method from those obtained by continuous recording. Similarly, their results indicated that whole-interval methods underestimated and partial-interval methods overestimated the actual duration of the measured behaviors, whereas momentary time-sampling produced the least distortion. Effects were most notable for medium duration behaviors across all coding systems, followed by brief duration behaviors and long duration behaviors. The authors suggested that momentary time-sampling methods should be used unless the minimum duration of the target behavior is less than the duration of the observation interval, at which point partial- or whole-interval recording methods may be satisfactory.

The results of these studies have demonstrated that the reliability of data differs depending on what time-sampling method is used. Specifically, across the studies examined, partial-interval methods tended to overestimate rates of behavior whereas whole-interval methods tended to underestimate behavior. In contrast, momentary time-sampling methods generally produced the most reliable estimates of behavior when compared to continuous samples. The duration of a target behavior was also shown to influence the amount of error associated with each system. Although these studies provide some insight into how reliability is influenced by time-sampling method, the targets of these studies (i.e., adult secretaries and children with intellectual impairments) and behaviors of interest (i.e., in-seat time and
stereotyped behaviors) are difficult to generalize to more common classroom behaviors and targets. More research is needed to see if these findings hold true for more typical classroom-based behaviors (e.g., academic engagement).

**Interval Length**

Interval length is another important consideration when developing a systematic direct observation coding system. When using time-sampling methods, interval length defines either how long a behavior has to last to be coded (i.e., in the case of whole-interval coding) or how frequently a behavior can be coded (i.e., in the case of partial-interval coding). In the studies above, interval length was shown to have an effect on the reliability of time-sampling methods, with shorter interval lengths yielding more reliable results. A few studies have specifically investigated this phenomenon to date.

As discussed, Powell and colleagues (1975, 1977) used varying interval lengths to examine differences in the reliability of time-sampling methods. In the original study conditions, where in-seat behavior occurred at steady rates, intervals of 120-, 80-, and 40-seconds were found to yield reliable results using momentary time sampling, partial-interval, and whole-interval methods, respectively (Powell et al., 1975). When behavior rates were manipulated in their follow-up study, however, shorter intervals were needed to obtain reliable data. For example, 60-second intervals were necessary when using a momentary time-sampling method, which is half the length found in the original study conditions (Powell et al., 1977). These findings suggest that variable behaviors may be more sensitive to changes in interval length than behaviors that occur more steadily.

Murphy and Goodall (1980) also manipulated interval length while investigating the reliability of different time-sampling methods. Specifically, data were collected using both 2.5-
second and 10-second partial interval coding. Using a 2.5-second partial-interval coding system produced more accurate results when compared to 10-second intervals for the targeted behaviors, also suggesting that shorter intervals may be preferable when possible.

A study by Sanson-Fisher, Poole, and Dunn (1980) specifically investigated the role interval length plays in determining the reliability of direct observation data. Eight patients from an acute short-stay psychiatric unit were observed using video recording methods. Each patient was observed on eight separate 5-minute occasions over 4 consecutive weekdays for seven categories of behavior (i.e., on task, talk, bizarre, etc.). Using real time recording equipment, a computer was used to analyze the data at different interval lengths ranging from 1 second to 10 seconds. If more than one behavior occurred within each interval, the behavior that occurred for a larger portion of time was coded. The number of occurrences of each behavior and the total time the subject was engaged in each behavior were coded using each interval length. These data were then compared to those of the 1-second interval data, as this was considered to be the most reliable dataset (Sanson-Fisher et al., 1980). Increasing the interval length from one second incrementally decreased the number of occurrences of a behavior for each target. In contrast, there was a tendency for the average duration of a behavior to increase as the interval length increased. The authors concluded that choice of interval length should be based on the duration of the target behavior, with shorter interval lengths necessary for behaviors that are brief in duration.

Saudargas and Zanolli (1990) investigated the influence of interval length by comparing the results of momentary time sampling against estimates obtained through real time recording. Sixteen elementary school children were observed during independent seatwork. Two children were observed on three occasions, two were observed on two occasions, and the remaining 12
children were observed once. The SECOS (Saudargas & Fellers, 1986), a multi-behavior observation system, was used to code student behavior over a 20-minute period. Although the SECOS can also be used to observe 12 event behaviors, only the six state behaviors (i.e., schoolwork, looking, other activity, child interaction, teacher interaction, and out of seat) were analyzed in the study. Two versions of the SECOS were used: a) the pencil and paper version (SECOS-MTS), which used 15-sec momentary time sampling to code the target behaviors, and b) a handheld computer version (SECOS-C), which continuously coded the target behaviors. When comparing the results obtained by each version, the authors found that 15-sec momentary time sampling closely approximated continuously coded data for most behaviors (i.e., other activity, child interaction, teacher interaction, and out of seat) behaviors (i.e., other activity, child interaction, teacher interaction, and out of seat) behaviors (i.e., other activity, child interaction, teacher interaction, and out of seat). When behaviors were short in duration, the authors cautioned that 15-sec intervals may not produce accurate estimates of behavior, and recommended using shorter interval lengths to obtain more accurate results.

In studies that have manipulated interval length to inform the reliability of systematic direct observation data, results have been consistent: increased interval length has been shown to demonstrate increased levels of error. When behaviors are long in duration, longer interval lengths may be acceptable. However, for short duration behaviors, shorter intervals may be necessary to detect instances of behavior change. Concerns can be raised, however, regarding how these results may generalize to more common classroom concerns given the targets and subjects in addition to the interval lengths explored by these studies, which are not consistent with those typically used in school based settings and coding schemes (i.e., 15-30 seconds).

**Observation Length**
Another area that has not been given much attention but is also important to consider when designing a systematic direct observation coding system is how long each observation session will last. As observation length increases, the number of response opportunities for collecting data also increases. The impact that increased observation length has on observational coding procedures has been explored in the studies below.

Karweit and Slavin (1982) also investigated the influence of observation length in their study of procedural factors influencing time-on-task data. The authors compared means and standard deviations of data from the first 10 minutes, first 20 minutes, and first 30 minutes of each observation session. Although it was difficult to predict how changes in observation length impacted the reliability of data due to variable results across classrooms, the authors generally concluded that obtaining shorter time segments of behavior appreciably altered the obtained results of a given observation.

Mudford, Beale, and Singh (1990) investigated the influence of observation length on the reliability of systematic direct observation data by observing five adults with intellectual impairments and physical handicaps in an institutional training setting. A real-time recording system utilizing one second momentary-time sampling was used to code behavior into six mutually exclusive and exhaustive categories (i.e., social interaction with peer, handling materials, stereotypy, etc.) over a two and a half hour period. Computer simulations were used to sample durations ranging from 15 to 135 minutes from the whole session records. A percentage similarity statistic was then computed by dividing the sampled observation data by the whole observation data and multiplying by 100, similar to a smaller/larger index (Hintze, 2005). The representativeness of the sampled durations, when compared to the whole-session records, varied based on (a) how long a subject engaged in a behavior once it was initiated and (b) the length of
the sampled observation session. High duration behaviors were estimated within 20% error by samples of less than 60 minutes, whereas low duration behaviors could not be accurately represented up through 135 minutes. These differences were attributed to the irregularity of low-duration behaviors, making them more sensitive to changes in observation duration. The authors concluded that the representativeness of observational data should not be assumed, and differences in reliability will take place based on subject, setting, and behavioral coding system. They recommended that adequate observation length should be empirically determined through exhaustive observations (Mudford et al., 1990).

McWilliam and Ware (1994) used G theory to examine the influence of the length of the observation sessions on reliability. Forty-seven preschoolers were observed in a university daycare setting, fifteen of whom were identified as having mild to moderate disabilities. Four types of engagement (with adults with peers with materials, and nonengaged) and five engagement levels (attentional, undifferentiated, differentiated, encoded, and symbolic) were targeted using a momentary-time sampling method involving 10-second intervals over a 15-minute period. Data were collected on four occasions. Analyses were conducting using the first 5 minutes, first 10 minutes, and all 15 minutes of each session to draw comparisons across observation lengths. Dependability coefficients were found to increase as the length of the observation increased. This pattern was consistent across both frequently and infrequently occurring behaviors, although the degree of improvement varied.

Most recently, Ferguson, Briesch, Volpe, and Daniels (2012) investigated the influence of observation length on the dependability of systematic direct observation data both within and across days using G theory. Twenty 7th grade students nominated by their teachers as exhibiting high levels of off-task behavior were observed during a regular education math activity. A 15-
second momentary time-sampling method was used to code for academic engagement. The greatest proportion of variance was attributable to the error term (50%). This was followed by variations across subjects (29%) as well as the relative standing of subjects across observations (15%). Dependability improved as a function of extending the length of each observation. Acceptable levels of dependability for progress monitoring were achieved after two 30-minute observations, three 15-minute observations, or four to five 10-minute observations when interested in estimating engagement across days. Even after five hour-long observations, acceptable levels of dependability for high-stakes decision-making could not be achieved. Within a given day, a 15-minute observation of academic engagement was determined to be adequate for low-stakes decision whereas an hour long observation session was necessary for high-stakes decisions.

These studies indicate that increasing observation length may yield higher levels of reliability for a variety of behaviors. Again, however, the degree of improvement for increased observation length varied by behavior category and the coding system being used. Behaviors that occur at moderate to steady rates may be captured reliably using short observation lengths, whereas variable or infrequent behaviors may require longer observation periods to be accurately sampled. Although some research has been conducted in this area that is relevant to school-based concerns, additional research must be conducted to confirm these findings as well as to investigate the influence of observation length on other classroom targets.

Number of Observations

Compared to other factors that affect the reliability of direct observation data, the number of observations needed to obtain reliable data has received the most attention. Increasing the number of observation sessions would increase the number of opportunities to collect data
points. The influence of number of observations on the reliability of systematic direct
observation data has been explored through both CTT and modern test theory approaches in the
studies below.

At the Annual Convention of the American Psychological Association, Marcus (1980)
presented research that used G theory to investigate the dependability of observational data. He
was interested in exploring the short term stability of observational data over a two month period
to provide guidelines around how large a behavior sample is necessary to produce dependable
data. The cooperative play of two samples of 31 preschoolers was examined over a two month
period. Six observations took place during free play activities in a nursery school setting using
10-second intervals over 24 minutes. A majority of the variance in estimates of cooperative play
was accounted for by error (76%) followed by differences between target students (23%). The
reliability of cooperative behavior in both samples was shown to be stable over time and
influenced by the overall amount of cooperative behavior demonstrated. Adequate levels of
dependability could not be achieved under the original study conditions ($E \hat{\rho}^2 \Phi = .68$), with
results demonstrating that a minimum of 10 observations of cooperative play ($E \hat{\rho}^2 \Phi = .81$)
would be needed in order to achieve an acceptable level of reliability. G theory proved a useful
tool for analyzing contributions to performance and estimating the size of a behavioral sample
that is necessary for a given target behavior (Marcus, 1980).

During their investigation of factors that influence the outcome of systematic observation
data for time-on-task, Karweit and Slavin (1982) also manipulated the number of days of
observation to see how this influenced reliability. This was accomplished by viewing each day of
observation as an item on a scale of total time-on-task (i.e., all 9 days of data) and calculating
internal consistency. Coefficient alphas were calculated to determine how consistent behaviors
were across differing numbers of days. Acceptable coefficient alphas were obtained subsequent to five days of observation (α = .71), whereas more desirable levels (i.e., .80 and above) were rendered after eight observation sessions (α = .81). The authors suggested that the number of observation sessions necessary to obtain reliable data may depend on the purpose of the evaluation, with more observations yielding higher levels of reliability.

While investigating the influence of observation length on the reliability of data, McWilliam and Ware (1994) also explored the optimal number of sessions necessary to reliably obtain general conclusions about a child’s engagement. The number of observations necessary to obtain reliable estimates of behavior varied depending on behavior type and level, with between 12 (for engagement with peers) and 40 (for engagement with materials) 15-minute observations necessary given a single observer. Even the most reliable behaviors required 3 hours of observation to obtain dependable results. The authors found that infrequently occurring behaviors were as reliable (or unreliable) as frequently occurring behaviors, and they questioned the reliability of studies that sampled behavior only once or twice to draw conclusions.

Doll and Elliot (1994) conducted an experiment to investigate how many occasions preschool children needed to be observed to accurately estimate their social behaviors. Twenty-four children attending a full-day laboratory preschool program at a major Midwestern university served as participants for the study. A total of nine 20-minute observations took place for each student using a 15-second partial-interval time-sampling method targeting 13 mutually exclusive categories of social behavior (e.g., Share, Affection, Ignore). Each observation was broken down into two 10-minute segments, which were analyzed in two ways. First, to examine how consistent data were from the beginning to the end of the observation session, data from the beginning and end of the session were compared, equivalent to test-retest analysis. Next, the
results of the complete data set were compared via correlational methods to those of subsequent increasing 10-minute periods (i.e., the first observation, the first two observations combined, the first three observations combined, and so on) to provide an estimate of internal consistency. This was done to provide insight into the number of observation sessions needed to be representative of the complete data set. Results indicated that reliability varied by behavioral category, and at least five observations were necessary to obtain reliable results for most of the behavior categories. None of the categories produced a reliable sample of social behavior with a single 10-minute observation.

McKevitt and Elliot (2005) also investigated the reliability of direct observation data across observation sessions used to describe the social behavior of preschoolers. The unstructured free play activities of two preschoolers were videotaped and coded using the Social Skills Rating System (SSRS; Gresham & Elliot, 1990) observation code designed to accompany the Social Skills Rating System – Teacher version (SSRS-T). Observation targets were determined based on SSIS-T results and appropriateness for observational methods, and event recording methods were used to determine the frequency of each behavior over the observation period. A total of twelve 30-minute observations were conducted. To assess reliability of the data across sessions, Pearson correlations were used to compare behavior frequencies from the first observation to the total 12, the second two observations sessions to the total 12, the first three observations to the total 12, and so forth. Statistical significance of the correlations was determined by comparing the observed correlations to their predicted value using Fisher’s $z$ transformations ($p \leq .05$; one-tailed). About three 30-minute observations were necessary to obtain reliable estimates of the social behavior of preschoolers ($r = .96 - 1.00$), demonstrating slightly more conservative estimates than Doll and Elliot (1994).
In their study described above, Hintze and Matthews (2004) also examined the number of observations needed to obtain dependable estimates of academic engagement. They found that for students with relatively stable levels of on-task/off-task behavior, only two or three observations were necessary. For those with variable levels of behaviors, however, estimates were much higher, with four observations per day across four school weeks deemed necessary to achieve adequate reliability. The authors cautioned that adequate levels of reliability may be difficult to achieve even with the simplest operational definitions and data collected over an extended period of time.

Volpe and colleagues (2009) also explored the influence of number of observations while investigating timing of observation. Results suggested that a total of 14 10-min observations would be needed to reach a dependability criterion of .80 for on-task behavior. They found that readily observable, overt behaviors (i.e., Hyperactivity-Impulsivity) required less observation when compared to behaviors that are less easy to observe (i.e., Sluggish Cognitive Tempo) (Volpe et al., 2009). The authors cautioned that conducting only one or two observations of student behavior may not be defensible practice, and the behavioral dimension of interest should be considered when determining how many observations to perform.

Most recently, Briesch, Chafouleas, and Riley-Tillman (2010) investigated the number of systematic direct observations necessary to obtain reliable academic engagement data while comparing these data to those collected through direct behavior rating. Video recordings of twelve kindergarten students in an inclusive classroom setting were coded for academic engagement using a 15-second momentary time-sampling method. Observations were approximately 10-15 minutes in length but ranged in correspondence with natural breaks in classroom activities. The highest proportion of variance in observation scores was attributable to
differences across subjects (48%), with error contributing a majority of the remaining variance (44%). Systematic direct observation data were found to be more dependable than DBR ratings, with acceptable levels of dependability achieved subsequent to three days of observation ($E \hat{\rho}^2 \Phi = .82$) whereas levels of reliability appropriate for making screening decisions were only achieved with DBR after 20 observations ($E \hat{\rho}^2 \Phi = .70$). These estimates were much more conservative than in previous studies. In situations where decisions need to be made quickly, the authors suggested that systematic direct observation may be a more attractive alternative to DBR ratings due to feasibility (Briesch et al., 2010). This is because of the higher level of precision that is associated with systematic direct observation methods when compared to DBR methods.

Although there have been many studies that have investigated the number of observation sessions needed to obtain reliable data, it is difficult to come to any universal conclusions. Studies varied widely by target behaviors and observational procedures, which have been shown to influence the reliability of data. However, this literature provides general insight into the influence of number of observations. These investigations found that reliability tends to improve as the number of observation sessions increases. However, the degree of improvement may vary considerably based on the target behavior, with more observations necessary to obtain reliable data for infrequent or variable behaviors. In terms of school based practice, several of these studies provided insight into how many observations may be necessary to obtain reliable estimates of student engagement (i.e. 3-5 15-minute observations), but results were also somewhat inconclusive, with reliability varying depending on procedures, subjects, and operational definitions. Across all of these studies, it is clear however that one observation may not be sufficient to provide a reliable estimate of behavior.

Conclusions and Implications for Research and Practice
Direct observation is commonly used by school psychologists to assess student behavior (Wilson & Reschly, 1996), and consideration of the psychometric properties of these data is important if the data are used to inform educational decision-making. Although research regarding the psychometric properties of systematic direct observation data is somewhat limited compared to other assessment methods, studies in this area have provided some insight into how different aspects of an observation procedure can affect the reliability of data. Both CTT and G theory have been used to explore these issues. The extent to which some of these findings can be generalized to school based practice, however, is questionable.

The effects of behavioral definition, time sampling method, interval length, observation length, time of observation, and number of observations on the reliability of systematic direct observation have been explored to varying degrees. Slight variations in operational definition were shown to produce different results when using systematic direct observation codes to measure on-task behavior (Karweit & Slavin, 1982). With regard to time of observation, time-on-task behavior may be generally consistent over some periods of time, but classroom activity may play an important role in the reliability of data (Hintze & Matthews, 2004; Karweit & Slavin; Volpe et al., 2009). Momentary time-sampling methods were found to produce the least amount of systematic error when compared to partial- and whole-interval methods, with these differences being more pronounced for behaviors that occurred infrequently or variably (Murphy & Goodall, 1980; Powell et al., 1975; Powell et al., 1977). The reliability of systematic direct observation data was generally found to increase as interval length was shortened (Murphy & Goodall, 1980; Powell et al. 1975; Powell et al., 1977; Sanson-Fisher et al., 1980) and as overall observation length increased (Ferguson et al., 2012; Karweit & Slavin, 1982; McWilliam & Ware, 1994; Mudford et al., 1990). Finally, reliability was found to improve as the number of
observation sessions increased (Briesch et al., 2010; Doll & Elliot, 1994; Hintze & Matthews, 2004; Karweit & Slavin, 1982; Marcus, 1980; McKeivitt & Elliot, 2005; McWilliam & Ware, 1994; Volpe et al., 2009). Overall, the degree to which each all of these factors impact the reliability of data was found to vary according to behavioral targets and their dimensions, with duration and frequency playing an important role.

In summary, there have been a relatively small number of studies conducted over the past four decades that have explored the psychometric properties of direct observation, and due to variations in behavioral targets, it is unknown whether many of these findings generalize to common school-based behaviors such as academic engagement. In terms of choosing a time-sampling method, momentary time-sampling methods were shown to consistently demonstrate the highest levels of reliability across all targets and coding schemes, suggesting that this likely holds for academic engagement as well. The impact of number of observations has received the largest amount of attention, and several of these studies investigated the common school-based target of student engagement (Briesch et al., 2010; Hintze & Matthews, 2004; Karweit & Slavin, 1982; Volpe et al., 2009). Although the impact of behavioral definition, observation length, and time of observation has also been examined under the lens of student engagement (Ferguson et al., in press; Hintze & Matthews, 2004; Karweit & Slavin, 1982; Volpe et al., 2009), more research is needed to support these findings. In addition, research in the area of interval length has only been conducted with subjects and behavioral targets that are difficult to generalize to school-based practice. Overall, research in the areas of behavioral definition, observation length, time of observation, and interval length as they relate to common school-based behaviors is needed to help inform best practices for practitioners using systematic direct observation.
References


Ferguson, T., Briesch, A., Volpe, R., & Daniels, B. (2012). The influence of observation length


CHAPTER II

Abstract

Direct observation is a tool commonly employed by school psychologists to investigate student behavior, and because these data are used for educational decision-making, accuracy is an important consideration. Procedural aspects of systematic direct observation have been shown to influence reliability, and this study was designed to explore how interval length and observation length influence the dependability of academic engagement data when using a momentary time sampling procedure. Twenty 7th grade students were each observed for two 30-minute sessions during math instruction. A series of generalizability (G) studies were conducted using combinations of common interval and observation lengths to examine how these factors influenced reliability-like coefficient. In general, shorter interval lengths and observation durations produced higher levels of dependability. Implications for practice and research, as well as limitations of the current study, are discussed.
EMPIRICAL RESEARCH PROPOSAL

Introduction

Although there are a number of methods that can be employed to investigate student behavior, direct observation is considered a foundational technique in the area of behavioral assessment (Landau & Swerdlik, 2005). This may be related to the fact that observation provides the most direct assessment of a student's behavior (Cone, 1977), which produces data that are more objective and less biased than those collected through indirect assessment methods (Merrell, 2008). There are many advantages to using direct observation, including: (a) noting the frequency of a behavior in a natural environment, (b) systematically examining the antecedents and consequences of a behavior, (c) informing the development of hypotheses regarding behavioral function, and (d) providing insight into the function of a behavior instead of focusing on topographical and descriptive accounts (Hintze, Volpe, Shapiro, 2002). Due to the utility of this method, direct observation has been identified as a vital component of the assessment of emotional and behavioral disorders as well as the functional behavioral assessment process (McConaughy & Ritter, 2008; Shapiro & Heick, 2004; Steege & Watson, 2008). Indeed, direct observation methods are popular among school psychologists, with practitioners reporting that they conduct 10 to 15 direct observations over the course of a typical month (Stinnett, Havey, & Oehler-Stinnett, 1994; Wilson & Reschly, 1996).

Although practitioners report using naturalistic observation nearly twice as often as systematic observation techniques (Wilson & Reschly, 1996), these data offer no empirical support for decision making processes because they are purely descriptive in nature (Hintze et al., 2002). In contrast, systematic direct observation provides quantitative descriptions of behavior using standardized, predetermined data collection procedures, which provide data that
can be used to inform decision making (Hintze et al., 2002). Target behaviors are operationally defined and coding systems are developed before the observation takes place, allowing for behavior to be compared across different times and settings (Salvia, Ysseldyke, & Bolt, 2010). Data collected through systematic direct observation can be used for a number of purposes, including traditional diagnostic assessment as well as intervention planning and monitoring.

One of the advantages of systematic direct observation is the inherent flexibility of this approach (Briesch & Volpe, 2007). Direct observation can be used to assess a wide range of observable behaviors, and procedures can be structured in a number of ways to collect information on different aspects of a behavior. Event recording methods can be used to describe the frequency and rate of discrete behaviors with a clear onset and offset (e.g., hand raises, call-outs) (Merrell, 2008). On the other hand, state behaviors (e.g., academic engagement, tantruming), which are prolonged conditions as opposed to specific, salient events, can be investigated using a few different data recording methods, including duration recording (i.e., the amount of time a behavior lasts) and latency recording (i.e., the amount of time that elapses before a behavior is initiated). When interested in describing behaviors that occur moderately to steadily, however, time-sampling methods (i.e., dividing the observation period into equal intervals and recording whether or not a specific target behavior occurs within each interval) are most appropriate (Hintze et al., 2002). For example, academic engagement is a popular observational target due to the important implications this behavior has in the learning process (Greenwood, 1996), and rates of this behavior are best captured through time sampling methods. Research has shown that when using time-sampling methods, momentary time-sampling techniques (i.e., recording the occurrence or non-occurrence of a behavior at the moment an interval begins or ends) may produce more accurate results when compared to partial-interval
INFLUENCE OF INTERVAL AND OBSERVATION LENGTH

(i.e., recording the occurrence of a target if it occurs at any time during an interval) and whole-interval (i.e., recording the occurrence of a behavior only if it occurs for the entire duration of an interval) techniques, which tend to overestimate or underestimate rates of behavior, respectively (Ary, 1984).

Although the flexibility of systematic direct observation procedures certainly make this a useful assessment approach, one of the challenges is in trying to evaluate the psychometric adequacy of direct observation data. Because of the significant impact that some educational decisions may have on students (e.g., special education eligibility, classroom placement), it is important that direct observation data are reliable and valid; however, the psychometric properties of systematic direct observation have not received as much attention as many other assessment methods (Leff & Lakin, 2005; Volpe, DiPerna, Hintze, & Shapiro, 2005). This may be related to the fact that there are essentially limitless possibilities in designing a systematic direct observation code. Whereas reliability and validity are expected to be fairly stable across testing situations when standardized procedures and instruments are used (e.g., standardized cognitive or achievement tests), there does not exist a single standardized direct observation approach. The flexibility inherent to systematic direct observation measures therefore becomes problematic for determining technical adequacy in that separate evaluations would be needed to investigate each unique observational procedure.

Because the reliability and validity of an assessment instrument are tied directly to the testing conditions (i.e., environmental, situational, and procedural factors), research investigating the psychometric properties of systematic direct observation data should therefore be considered in the context of providing insight into how different procedural aspects affect reliability and validity to inform general guidelines for implementation. When examined through mathematical
formulas, research suggests that using momentary time-sampling procedures over partial- and whole-interval coding systems may minimize systematic error (Ary, 1984). Although there are a number of other aspects that may influence the technical adequacy of systematic direct observation coding systems using momentary time-sampling methods (e.g., behavioral definition, timing of the observation, number of observations), two important considerations of procedural design that hold regardless of behavioral target are the length of the individual intervals and the length of the overall observation.

**Interval length**

Interval length is an important consideration, as it determines how many response opportunities will occur over the duration of the observation session. For example, assuming a 15-min observation, a student’s behavior would be assessed 60 times using 15-sec sampling versus only 30 times using 30-sec sampling. From a traditional assessment perspective, increasing the amount of response items is a common way to improve reliability because chance errors of measurement begin to cancel out as the number of response opportunities increases (Thorndike, 2005). Studies that have explored the impact of interval length on the reliability of systematic direct observation data have relied on comparing data obtained using different interval lengths to those obtained through continuous coding procedures. Powell, Martindale, and Kulp (1975) used momentary time sampling to examine the in-seat behavior of a secretary (i.e., posterior is in contact with the seat) at 10-, 20-, 40-, 80-, 120-, 240-, 400- and 600-second intervals. In-seat behavior occurred at steady rates over the course of the observation sessions, and intervals of up to 120 seconds were found to reliably describe in-seat behavior. However, when behavior was engineered to demonstrate more variability in a follow up study, Powell, Martindale, Kulp, Martindale, and Bauman (1977) found that reliable data could only be
collected using intervals of up to 60 seconds, half the length found under the original study conditions. Saudargas and Zanolli (1990) compared 15-sec momentary time sampling to continuous coding when using the SECOS to observe 16 elementary school students during independent seatwork. They found that most behaviors (i.e., other activity, child interaction, teacher interaction, and out of seat) could be accurately represented using 15-sec momentary time sampling, but suggested that shorter interval lengths may be needed for behaviors that are short in duration.

Two additional studies investigating the influence of interval length examined much shorter intervals. When Murphy and Goodall (1980) used partial-interval time-sampling methods to describe three brief duration (i.e., fist to chin, pinching self, and vocalization), medium duration (i.e., hand flapping, body and hand tapping/posturing), and long duration (i.e., body rocking, finger sucking, and hand posturing) behaviors of eight intellectually impaired children, they found that using a 2.5-second interval coding system produced more accurate results when compared to a 10-second interval coding system. Sanson-Fisher, Roole, and Dunn (1980) observed eight patients from an acute short-stay psychiatric unit for seven behavior categories (e.g., on task, talk, bizarre), and a computer was used to generate data sets based on different interval lengths ranging from 1 second to 10 seconds. When increasing interval length, the number of occurrences of a behavior for each target was found to decrease whereas the average duration of a behavior was found to increase.

Studies that have investigated the influence of interval length on the reliability of systematic direct observation data suggest that increasing interval length generally increases levels of error. Although longer interval lengths may be acceptable for describing behaviors that are long in duration, shorter intervals may be necessary to detect instances of behavior change.
for short duration behaviors. Given both the targets and subjects of these studies, however, it is unclear whether these findings would generalize to typical school-based concerns.

**Observation length**

The length of the observation session is also important to consider when developing a momentary time-sampling system as this will also determine the number of response opportunities when collecting data. Because more data can be collected over larger periods of time, it seems likely that longer observation sessions would result in stronger reliability as error begins to cancel out. Karweit and Slavin (1982) compared means and standard deviations of time-on-task data from the first 10-minutes, first 20 minutes, and first 30 minutes of the observation sessions. Although results varied across classrooms, the authors generally concluded that obtaining smaller samples of behavior appreciably altered the obtained results of a given observation. Mudford, Beale, and Singh (1990) used a real-time recording system to code the behavior of five adults with intellectual impairments and physical handicaps into six mutually exclusive and exhaustive categories (e.g., social interaction with peer, handling materials, stereotypy). Computer simulations were used to sample durations ranging from 15 to 135 minutes from the whole observation records. The amount of time a subject engaged in a behavior and the length of the sampled observation session were found to influence the representativeness of scores when compared to the complete observation record, with low duration behaviors and shorter sampled observation sessions yielding less accurate results.

Two studies have used generalizability (G) theory to investigate the influence of observation length on dependability. McWilliam and Ware (1994) compared the reliability-like coefficients of data describing the free play of preschoolers collected during the first 5 minutes, first 10 minutes, and the full 15 minutes an observation period. Results suggested that the
minimum number of observation sessions needed to achieve a dependability level of .80
decreased as the length of the observation increased. Recently, Ferguson, Briesch, Volpe, and
Daniels (2012) investigated the influence of observation length on the dependability of
systematic direct observation using G theory. A 15-second momentary time-sampling method
was used to code the academic engagement of 20 7th grade students nominated by their teacher
as exhibiting high levels of off-task behavior. Dependability was found to improve as the
duration of each observation session increased. Acceptable levels of dependability for progress
monitoring were achieved after two 30-minute observations, three 15-minute observations, or
four to five 10-minute observations when interested in describing behavior across days. Within a
given day, a 15-minute observation was determined to be adequate for low-stakes decision
whereas an hour long observation session was necessary for high-stakes decisions.

Increasing the length of the observation session has been shown to yield higher levels of
reliability. However, similar to interval length, the degree of improvement varied by the
behavioral category and coding system being used. When behaviors occur at moderate to steady
rates, shorter observation sessions can be used to obtain reliable data. Comparatively, variable or
infrequent behaviors require longer observations sessions to yield reliable estimates. Although
some of this research relates to school-based practice, additional research is needed to confirm
and expand upon these findings.

**Purpose of Study**

Although systematic direct observation is an assessment technique commonly used to assess
student behavior (Shapiro & Heick, 2004; Stinnett et al., 1994; Wilson & Reschly, 1996), and
using momentary time-sampling procedures produces the least systematic distortion in data (Ary,
1984), there are few guidelines available regarding the best way to use momentary time sampling
in school settings. Preliminary research suggests that reliable academic engagement data for the purpose of progress monitoring can be collected through two 30-minute, three 15-minute, or four to five 10-minute observations (Ferguson et al., 2012). However, additional research is necessary to determine if these findings hold true with different student populations. Furthermore, although interval length has also been shown to influence the reliability of direct observation data (Murphy & Goodall, 1980; Powell et al. 1975; Powell et al., 1977; Sanson-Fisher et al., 1980), the results of these studies are difficult to generalize to common school-based systematic direct observation targets, such as academic engagement. Therefore, the present study was designed to answer the following research questions:

1. Can the results obtained in Ferguson et al. (2012) be replicated with a new group of students? That is:
   a. Does the recommendation of two 30-min observations, three 15-min observations, or four to five 10-min observations to achieve an acceptable level of dependability across day hold across students?

2. Are there substantial differences (i.e., percent difference of 5% or greater) in the proportion of score variance attributable to particular facets (i.e. person, occasion, error) depending on the interval length selected?

3. Do substantial differences exist in the duration and number of occasions needed in order to obtain a dependable estimate of student engagement based on the interval length that is selected?

Method

Participants
Study participants were students enrolled in one of four sections of an academic class taught by the same classroom teacher. Students were in the 7th grade, predominantly of African American and Latino descent, and attended an urban charter school in the Northeastern United States. In light of recommendations from Webb, Rowley and Shavelson (1988) indicating that G studies include a minimum of 20 persons and 2 conditions per facet, observations were conducted using a total of 20 students. The classroom teacher was provided with a definition of academic engagement and asked to nominate at least five students in each section who exhibited low levels of academic engagement. Students who were clearly visible from at least one camera angle and present for both observation sessions were selected from this pool to participate in the study. Because students typically observed in schools are those exhibiting behavioral problems, those nominated by their teacher as exhibiting lower levels of academic engagement were more likely to approximate students that are generally observed by practitioners (Volpe, McConaughy, & Hintze, 2009).

**Dependent Variable**

Academic engagement was targeted using a modified version of the definition used in the Behavioral Observation of Students in Schools (BOSS; Shapiro, 2004). Similar to Hintze and Matthews (2004), Briesch, Chafouleas, and Riley-Tillman (2010), and Ferguson and colleagues (2012), active and passive engagement categories were combined into one category. *Actively engaged* behaviors were defined as any time the student was actively interacting with the academic environment (e.g. completing a worksheet, reading aloud, or answering a question). *Passive engagement* incorporated less overt types of engagement, such as silent reading, looking at the board, or listening to directions. Non-examples of active and passive engagement included
looking out the window, talking to a peer, or playing with objects. If a student was either actively or passively engaged, he or she was coded as academically engaged.

**Procedure**

**Systematic direct observation.** Data were collected using existing classroom video footage. Use of video recordings allowed observers to code the behavior of multiple students during the same observation period. Research has shown that observational data collected via video recording methods produces minimal systematic distortion compared to data collected through live observations (Kent, O'Leary, Dietz, & Diament, 1979). Two 30-minute observations across two class sessions were conducted, consistent with procedures used by Ferguson and colleagues (2012). The observation period was broken down into 5-minute blocks to contrast meaningful differences between adequate behavior samples based on different observation durations, similar to previous studies that have used G theory to investigate the influence of observation length (Ferguson et al., 2012; McWilliam & Ware, 1994). Roughly half of the observation session (i.e., blocks 1, 3 and 6) consisted of large group review of a math worksheet involving active questioning and participation between the students and instructor. The rest of the observation session consisted of independent seatwork (i.e., blocks 2 and 4) and small group work (block 5).

A momentary time-sampling method utilizing 5-sec intervals was used to collect data. A 5-sec beep track was used to code data by marking a “+” (academically engaged) or “-” (not academically engaged) for each interval on a recording sheet when the tone sounded. Although longer interval lengths (e.g., 15-sec, 30-sec) are more common in practice, using 5-sec intervals was necessary in order to answer research questions regarding the influence of interval length. Data for the other interval length conditions were calculated by taking the data collected with 5-
second intervals and using every second data point to generate data based on 10-second momentary time-sampling, every third interval to generate data based on 15-second intervals, and so forth.

Observer training. Data were collected by two graduate students in school psychology who had previous coursework and fieldwork experience using direct observation systems, including the BOSS (Shapiro, 2004). Training sessions were completed using the coding procedures described above to ensure that academic engagement was coded consistently across observers to a criterion of at least 80% agreement. The academic engagement of three students was observed using 10-minute training videos. Interobserver agreement (IOA) was determined by dividing the number of interval-by-interval agreements by the total number of interval and multiplying by 100 to obtain a percentage. IOA ranged from 88% to 96%, with an average of 92%. The corresponding kappa value was .74, thus indicating good agreement (Watkins & Pacheco, 2000).

Interobserver agreement. One observer served as the primary data collector and the second observer coded 35% of the data in order to assess reliability of the observational data. Percent agreement ranged from 84% to 98%, with an average of 95% and corresponding Kappa Value of .60, indicating a good level of agreement (Watkins & Pacheco, 2000).

Data Analysis

This study used G theory to examine the proportion of variance attributable to relevant measurement facets. The goal of G theory is to determine the extent to which a given set of measurements generalize to actual values of a given target (Cone, 1977). G theory provides insight into the dependability of a measure, which incorporates both reliability (i.e., how consistent behaviors are under specified conditions) and validity (i.e., extent to which the
observation reflects actual behavior over all possible conditions). First, the proportion of error associated with a situational variable relevant to the observation session is determined through the estimation of variance components within a G study (Cone, 1977). This information can be used to improve assessment procedures by identifying major sources of measurement error. The variance components obtained through G studies can then be used to simulate Decision (D) studies, wherein facets can be manipulated to derive reliability-like coefficients. For example, it is possible to estimate the number of observation sessions needed to obtain an adequate level of dependability. Dependability coefficients are used to evaluate absolute (i.e., within-individual) decisions, such as formative assessments, which do not take the performance of others into account. On the other hand, generalizability coefficients are used when making relative (i.e., between-individual) evaluations, such as when using norm referenced assessments (Shavelson & Webb, 1991). Discussion of results is limited to dependability coefficients because systematic direct observation data are most typically used to make absolute (i.e. intraindividual) rather than relative (i.e. rank-order) decisions.

Within the current study, the goal was to investigate three facets of interest: person, occasion, and block. Person was defined as each individual student who was observed in the classroom. Occasion was defined as each day on which an observation was conducted. Finally, block was defined as the estimate of academic engagement obtained within each 5 minute period of observation. Although interval length was also of interest, the intent was to run separate models for each interval length condition in order to draw meaningful comparisons across these interval lengths with regard to variance components and reliability-like coefficients.

For each interval length condition, the intended design was one in which persons were fully crossed with blocks and occasions (i.e. \( p \times b \times o \)). This means that every student was rated
during each 5-minute period across both observation sessions. This made it possible to estimate seven variance components: person, block, occasion, person x block, person x occasion, block x occasion, and a residual error term, which included the three-way interaction between persons, block, and occasion. Because the goal was to generalize the study results beyond the particular blocks and occasions sampled, these facets were treated as random. Variance components were derived in SPSS 19.0 using ANOVA with Type III Sum of Squares. These variance components were meant to be used in a series of subsequent decision (D) studies to determine how the dependability of scores would change as a result of changes in the length of the observation sessions and number of occasions.

**Results**

Students were observed for 30 minutes on two separate occasions for levels of academic engagement across 6 5-min blocks using a 5-sec momentary time sampling procedure. Minimal overall differences were found across occasions, with the mean levels of academic engagement across all 20 students ranging from 89.38% on the first occasion to 88.89% on the second occasion. Overall standard deviations also varied slightly, from 23.50 on the first occasion to 24.71 on the second occasion. Across 5-min blocks, rates of academic engagement were more variable. The level of academic engagement observed within a 5-min block ranged from 43.33% to 100% on the first occasion and 41.67% to 100% on the second occasion. When examining variability across all 6 5-minute blocks for a given student, differences in rates of academic engagement ranged from 3.33% to 46.67%, with an average difference across blocks of 22.25%. Variability across 5-min blocks for each student was slight higher on the second occasion, ranging from 0% to 55% with an average of 23.50%. See Table 1 for descriptive information.
Initial full model (i.e., \( p \times b \times o \)) G studies were conducted within each interval length condition; however, large negative variance components were found for the person x blocks interaction term under most interval length conditions. Large negative variance components indicate a misspecification of the model (Shavelson & Webb, 1991). Upon further examination, it was noted that levels of academic engagement for each 5-min block appeared to change substantially for some students over the course of the observation session, suggesting that specific instances of the facet of block (e.g., the first five minutes, next five minutes) were not interchangeable with one another. These discrepancies for some students may have diminished the capacity of this facet to provide meaningful information to the model. The decision was therefore made to remove the facet of block from the model and conduct separate G studies for common combinations of observation lengths (i.e. 10-min, 15-min, 20-min, 30-min) and interval lengths (i.e., 10-sec, 15-sec, 20-sec, 30-sec). The design used was one in which persons were fully crossed with occasions (i.e. \( p \times o \)). This means that every student was rated across both observation sessions. This made it possible to estimate three variance components: person, occasion, and a residual error term.

Data from each interval and observation length condition were first analyzed to determine the percentage of variance attributable to the measurement facets under the full model (i.e., persons, duration, and occasion) (see Table 2). Across roughly half \( (n = 9) \) of the conditions, the largest proportion of variance was attributable to differences in levels of academic engagement across students (i.e., person), the highest being 75% under the 10-sec, 10-min condition. For the remaining conditions, the residual error term contributed the highest proportion of variance, contributing up to 63% under the 30-sec, 15-min condition. Differences in academic engagement
across occasions (i.e., occasions) contributed negligible variance under most conditions, with the highest proportion being 5% under the 15-sec, 10-min condition.

A few patterns emerged regarding the percentage of variance attributable to the facet of person across conditions (see Figure 1). Changes regarding the percentage of variance attributable to person across interval length conditions were most clear under the 10-min condition, with a large and consistent decrease as interval length increased. A similar pattern existed for 15-min observations, with 10-sec and 15-sec intervals producing an appreciably higher percentage of variance for the facet of person when compared to the 20-sec and 30-sec intervals. At longer observation lengths (i.e., 20-min and 30-min), the pattern was less clear, though the highest percentage of variance for the facet of person was found when using 15-sec intervals. When looking within each interval condition, person explained the largest proportion of variance when conducting a 10-min observation, with a decrease in the size of this facet when conducting longer observation sessions. In general, the facet of occasion did not contribute meaningfully to the model regardless of interval or observation length.

Variance components were used to conduct a series of D studies in order to identify the conditions (i.e., length of observation, length of MTS interval, number of observations) under which a dependable estimate of academic engagement might be obtained. A level of dependability appropriate for progress monitoring or other low-stakes decision making purposes (i.e., .70 and above; Salvia, Ysseldyke, & Bolt, 2010) was obtained through a number of interval length and observation length conditions (see Table 3). Unlike Ferguson and colleagues (2012), dependable data could not be obtained through 2 30-min observations when using 15-sec intervals, with estimates being slightly lower ($\Phi = .68$ versus $\Phi = .70$). At shorter observation lengths, however, the current study produced higher reliability-like coefficients when conducting
3 15-min (Φ = .80 versus Φ = .71) or four (Φ = .90 versus Φ = .72) to 5 (Φ = .92 versus Φ = .76) 10-min observations.

Some general patterns were also identified in the obtained dependability coefficients. First, regardless of the interval length or observation length being used, dependability coefficients improved as the number of occasions increased. A point of diminishing returns occurred after 4 occasions, with small gains made after 5 occasions. Second, holding the number of observations constant, dependability coefficients generally decreased as the length of the observation session increased. Ten-minute observation sessions produced the most dependable data across all interval length and occasion combinations. In contrast, in nearly all cases, conducting 30-min observation sessions produced the least dependable data. Finally, regarding interval length, dependability coefficients consistently decreased as interval length increased under the 10-min condition, shorter interval lengths (i.e., 10 seconds and 15 seconds) produced higher dependability compared to longer interval lengths (i.e., 20 seconds and 30 seconds) under the 15-min condition, and the pattern was not clear for 20-min and 30-min observation lengths, though the 15-sec condition produced the most dependable data at these observation lengths.

**Discussion**

Previous research has shown that both interval length (Murphy & Goodall, 1980; Powell et al. 1975; Powell et al., 1977; Sanson-Fisher et al., 1980) and the overall length of the observation session (Ferguson et al., 2012; Karweit & Slavin, 1982; McWilliam & Ware, 1994; Mudford, Beale, & Singh, 1990) can influence the reliability of direct observation data. Specifically, the reliability of time-sampling data has been shown to increase with the use of shorter intervals and longer overall observation sessions. The current study was designed to validate previous findings regarding the influence of observation length on the dependability of
academic engagement data while also examining the influence of interval length on data outcomes when using a momentary time sampling procedure.

Whereas Ferguson and colleagues (2012) found that dependability improved as the length of the overall observation increased when using a momentary time sampling procedure to examine levels of academic engagement, the opposite held true for the current study; dependability generally decreased as the duration of the observation session increased. Ten minute observation sessions produced the most dependable data across all conditions, and 30-minute observation sessions generally produced the least dependable data. Differences in the patterns of academic engagement data across these studies may account for this discrepancy. In Ferguson et al. (2012), student behavior was relatively stable over the duration of the observation session, meaning that levels of academic engagement remained fairly consistent over time (i.e., academic engagement during the first five minutes was similar to the next five minutes). Therefore, as observation length increased, confidence in the ability to generalize from the obtained data to the students’ overall level of engagement for a time period increased as well. In the current study, however, levels of academic engagement fluctuated for some students as the length of the observation session increased (i.e., academic engagement varied considerably from one 5-minute segment to the next; see Table 1). Therefore, as additional 5-minute blocks of observation were added, the ability to generalize from the data to the students’ overall level of engagement for the time period actually decreased.

A parallel situation can be explained using raters as a facet of measurement. If a pool of observers were highly trained and rated very consistently, our confidence in the dependability of the data would increase with additional observers. For example, if two raters obtained a score of 80%, we would be fairly confident that their ratings were dependable. If five raters all obtained
80%, we would be even more confident in the obtained values. However, if observers received
minimal training and produced discrepant ratings, the opposite would hold true; adding
additional observers would make us even less confident that the obtained results are correct. For
example, if one observer obtains a rating of 80% and a second observer obtains a rating of 70%,
the true estimate of a behavior would be unclear. However, if the first observer rates 80%, the
second rates 60%, the third rates 70%, the fourth rates 50%, and so on, the true rate of a behavior
becomes even less clear.

Although the facet of observation duration (i.e., 5-minute blocks) explained meaningful
differences in academic engagement for the G studies completed by Ferguson and colleagues
(2012; blocks = 0%, person x blocks = 15%), substantial variability in student academic
engagement across 5-minute blocks in the current study meant that block could not be included
as a facet of measurement. Therefore, G and D study results could not be directly compared.
Generally speaking, however, the current study produced more dependable results at shorter
observation durations whereas Ferguson and colleagues (2012) obtained more dependable results
for longer observation sessions. For example, dependability coefficients were higher in Ferguson
and colleagues (2012; Φ = .70) than the current study (Φ = .68) when comparing two 30-minute
observations when using the same interval length (i.e., 15-seconds). However, estimates in the
current study (Φ = .80) were higher than previous estimates (Ferguson et al., 2012; Φ = .71)
when comparing three 15-minute observations.

The results of the current study both compare and contrast with other investigations of the
influence of observation length. Similar to Ferguson and colleagues (2012), McWilliam and
Ware (1994) found that reliability-like coefficients improved as the observation length increased
when observing the free play of preschoolers. However, improvements in dependability were
minimal for most behaviors when increasing from 10 to 15 minutes, and it is difficult to
determine how results would generalize to longer observation lengths such as those in the current
study. Mudford, Beale, and Singh (1990) also found that shorter observation lengths yielded less
representative data, but differences in methods (i.e., sampled durations from the whole
observation record) and observational targets (i.e., behaviors of adults with intellectual and
physical disabilities) make results difficult to compare.

During their investigation of factors influencing time-on-task data, Karweit and Slavin
(1982) also discussed the importance of observation timing as it relates to data outcomes. When
comparing rates of time-on-task during the first 10 minutes, the second 10 minutes, and the third
10 minutes of instruction in a mathematics class, they found that average time on-task varied
markedly across these time periods for some classrooms. They concluded that timing of
observation was important when rates of time-on-task are not evenly distributed, and the
selection of the observation period may be very consequential for data outcomes. In the current
study, classroom activity alternated between interactive whole group instruction and
independent/small group work across 5-minute blocks. Although not intended, this caused the
facet of block to essentially be the same as setting due to the changes that occurred in terms of
instructional activity across 5-minute time segments. In light of Karweit and Slavin’s (1982)
findings, it is possible that these changes in classroom activity appreciably altered rates of
student academic engagement over the course of the observation period, lowering the
dependability of longer observation sessions.

Because shorter interval lengths may minimize error by collecting more response items
(Thorndike, 2005) and have been shown to result in increased accuracy (Murphy & Goodall,
1980; Powell et al. 1975; Powell et al., 1977; Sanson-Fisher et al., 1980), a second goal of the
current study was to determine whether both variance component estimates and dependability coefficients differed depending on the interval length selected when observing for rates of academic engagement. Substantial differences existed in the percentage of variance attributable to the measurement facets of person and the error term based on the interval length being selected. In general, more of the variance was explained by differences across persons when using shorter intervals (i.e., 10 second and 15 second) compared to longer intervals (i.e., 20 seconds and 30 seconds; see Figure 1). The error term was generally the highest when using 30-sec intervals and lowest when using 10-sec intervals. Substantial differences in dependability coefficients also existed depending on the interval length being used. Typically, using shorter interval lengths (i.e., 10 second and 15 second) resulted in higher levels of dependability compared to longer interval lengths (i.e., 20 seconds and 30 seconds; see Table 3) when collecting academic engagement data, requiring few observation sessions to obtain dependable data when observation length was held constant.

**Limitations**

There are limitations to the current study that should be considered. All academic engagement data were extrapolated from a 5-second interval dataset. It is possible that data may have been different if actually collected using each interval length included in the study (i.e., 10-seconds, 15-seconds, 20-seconds, and 30-seconds). The results of the current study are also specific to the conditions under which the data were collected, as is the case with other studies investigating the psychometric properties of systematic direct observation data. It is unclear how these results may generalize to other observation targets in other settings. Due to feasibility concerns regarding data collection, data were only collected over two occasions. This is the minimum number of occasions needed to conduct the G study, and research has shown that
variance components may become more stable as instances of a facet increase (Smith, 1981). It is worth noting that Kappa Values for interobserver agreement fall at the bottom end of the “good agreement” category (Watkins & Pacheco, 2000), despite average percent agreement of 95%. Finally, data were collected through video tapes rather than in vivo. It may have been easier for the observers in the study to examine the student’s behavior than is typical for most practitioners collecting data in naturalistic settings. Therefore, dependability estimates may be higher than what may be found in general practice.

Implications for Research and Practice

The results of this investigation have implications for both researchers and practitioners who use systematic direct observation to collect data regarding student academic engagement using momentary time sampling procedures. This study demonstrates that rates of student behavior are important to consider when designing systematic direct observation system procedures. Previous research has shown that when academic engagement is relatively stable over time, dependability increases as observation length increases (Ferguson et al., 2012). In the current study, however, descriptive data indicates that rates of academic engagement were more variable, and increasing the length of the observation session subsequently lowered dependability. Classroom activity changed roughly every 5 minutes in the current study, which may have caused rates of academic engagement to shift over the course of the observation session. Previous research has demonstrated the importance of scheduling in reference to time-on-task data (Karweit & Slavin, 1982). The current study also speaks to the importance of scheduling the observation session during a consistent activity in order to limit the behavioral variability observed, and future systematic direct observation protocols should consider the types of instructional activities that occur over the course of the observation session. The activity
chosen should be the time during which the behavior in question is most likely to manifest to enhance the effectiveness of the observation session in determining the true rates of a behavior during a specific time period. Indeed, Shavelson, Webb, and Rowley (1989) recommended conducting an observation during the time period a target behavior is most likely to occur to enhance the effectiveness of the observation to describe true rates of a behavior during an activity of interest.

The current study also extends previous findings that suggest interval length influences the outcomes of systematic direct observation systems (Murphy & Goodall, 1980; Powell et al. 1975; Powell et al., 1977; Sanson-Fisher et al., 1980) in the context of a common school-based concern: academic engagement. This study found that 15-sec intervals generally produced similar or higher levels of dependability when compared to 10-sec intervals, suggesting that using intervals shorter than 15 seconds may not increase precision when collecting academic engagement data. This information may be useful for practitioners who utilize partial interval coding systems to target other behaviors while collecting momentary time sampling data on academic engagement. Using these systems may be cumbersome when using intervals smaller than 15-sec, but doing so would provide no appreciable benefits in terms of data outcomes. However, using longer interval lengths (i.e., 20 seconds and 30 seconds) was shown to noticeably lower the dependability of the data. Therefore, shorter interval lengths appear generally preferable to longer ones when collecting academic engagement data using momentary time sampling, but a point of diminishing returns may be reached as interval lengths become smaller.

Overall, the current study supports previous research concluding that small changes in procedural aspects of systematic direct observation may have a substantial influence on data
outcomes when using a momentary time sampling procedure to collect academic engagement data. Careful consideration should be given to the selection of interval and observation duration depending on the purpose of data collection. Additional research is needed to both confirm and expand upon the current findings. Because the results of this study conflict with one of the only other investigations of the influence of observation length on the dependability of academic engagement data, more research is needed to clarify how these results may generalize to other populations. Data for each interval length condition were extrapolated from a 5-sec interval database; future research may consider actually collecting data based on different interval lengths. Finally, future research may examine how these results generalize to other school based concerns, such as off-task behaviors.
INFLUENCE OF INTERVAL AND OBSERVATION LENGTH

References


INFLUENCE OF INTERVAL AND OBSERVATION LENGTH


### Table 1

**Mean (SD) Student Engagement using 5-sec Time Sampling by 5-min Block**

<table>
<thead>
<tr>
<th>Student</th>
<th>Occasion 1</th>
<th>Occasion 2</th>
<th>Occasion 3</th>
<th>Occasion 4</th>
<th>Occasion 5</th>
<th>Occasion 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Block 1: 0:05 – 5:00</td>
<td>Block 2: 5:05 – 10:00</td>
<td>Block 3: 10:05 – 15:00</td>
<td>Block 4: 15:05 – 20:00</td>
<td>Block 5: 20:05 – 25:00</td>
<td>Block 6: 25:05 – 30:00</td>
<td>0:05 – 30:00</td>
</tr>
<tr>
<td>1</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>95.00 (21.98)</td>
<td>78.33 (41.55)</td>
<td>56.67 (49.97)</td>
<td>100 (0)</td>
<td>88.33 (18.92)</td>
</tr>
<tr>
<td>2</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>96.67 (18.10)</td>
<td>91.67 (27.87)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>97.78 (9.81)</td>
</tr>
<tr>
<td>3</td>
<td>98.33 (12.91)</td>
<td>100 (0)</td>
<td>95.00 (21.98)</td>
<td>78.33 (41.55)</td>
<td>93.33 (25.15)</td>
<td>86.67 (34.28)</td>
<td>91.94 (22.64)</td>
</tr>
<tr>
<td>4</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>90.00 (30.25)</td>
<td>91.67 (27.87)</td>
<td>96.67 (18.10)</td>
<td>90.00 (30.25)</td>
<td>94.44 (19.90)</td>
</tr>
<tr>
<td>5</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>60.00 (49.40)</td>
<td>100 (0)</td>
<td>93.33 (8.23)</td>
</tr>
<tr>
<td>6</td>
<td>96.67 (18.10)</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>99.17 (5.17)</td>
</tr>
<tr>
<td>7</td>
<td>88.33 (32.37)</td>
<td>85.00 (36.01)</td>
<td>80.00 (40.34)</td>
<td>81.67 (39.02)</td>
<td>96.67 (18.10)</td>
<td>91.67 (27.87)</td>
<td>87.22 (32.29)</td>
</tr>
<tr>
<td>8</td>
<td>76.67 (42.65)</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>86.67 (34.28)</td>
<td>100.00 (0)</td>
<td>96.67 (18.10)</td>
<td>93.06 (17.99)</td>
</tr>
<tr>
<td>9</td>
<td>98.33 (12.91)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>81.67 (39.02)</td>
<td>86.67 (34.28)</td>
<td>94.17 (16.52)</td>
</tr>
<tr>
<td>10</td>
<td>86.67 (34.28)</td>
<td>100 (0)</td>
<td>96.67 (18.10)</td>
<td>93.33 (25.15)</td>
<td>96.67 (18.10)</td>
<td>78.33 (41.55)</td>
<td>91.94 (22.86)</td>
</tr>
<tr>
<td>11</td>
<td>70.00 (46.21)</td>
<td>75.00 (43.67)</td>
<td>43.33 (49.97)</td>
<td>80.00 (40.34)</td>
<td>90.00 (30.25)</td>
<td>83.33 (37.58)</td>
<td>73.61 (41.34)</td>
</tr>
<tr>
<td>12</td>
<td>65.00 (48.10)</td>
<td>76.67 (42.65)</td>
<td>51.67 (50.39)</td>
<td>50.00 (50.42)</td>
<td>83.33 (37.58)</td>
<td>65.00 (48.10)</td>
<td>65.28 (46.21)</td>
</tr>
<tr>
<td>13</td>
<td>96.67 (18.10)</td>
<td>83.33 (37.58)</td>
<td>81.67 (39.02)</td>
<td>90.00 (30.25)</td>
<td>75.00 (43.67)</td>
<td>86.67 (34.28)</td>
<td>85.56 (33.82)</td>
</tr>
<tr>
<td>14</td>
<td>61.67 (49.03)</td>
<td>88.33 (32.37)</td>
<td>81.67 (39.02)</td>
<td>86.67 (34.28)</td>
<td>93.33 (25.15)</td>
<td>88.33 (32.37)</td>
<td>83.33 (35.37)</td>
</tr>
<tr>
<td>15</td>
<td>91.67 (27.87)</td>
<td>96.67 (18.10)</td>
<td>83.33 (37.58)</td>
<td>76.67 (42.65)</td>
<td>73.33 (44.59)</td>
<td>85.00 (36.01)</td>
<td>84.44 (34.47)</td>
</tr>
<tr>
<td>16</td>
<td>91.67 (27.87)</td>
<td>98.33 (12.91)</td>
<td>91.67 (27.87)</td>
<td>96.67 (18.10)</td>
<td>93.33 (25.15)</td>
<td>95.00 (21.98)</td>
<td>90.00 (30.25)</td>
</tr>
<tr>
<td>17</td>
<td>96.67 (18.10)</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>96.67 (18.10)</td>
<td>95.00 (21.98)</td>
<td>91.67 (27.87)</td>
<td>96.39 (16.49)</td>
</tr>
<tr>
<td>18</td>
<td>88.33 (32.37)</td>
<td>98.33 (12.91)</td>
<td>71.67 (45.44)</td>
<td>95.00 (21.98)</td>
<td>73.33 (44.59)</td>
<td>100 (0)</td>
<td>87.78 (26.22)</td>
</tr>
<tr>
<td>19</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>95.00 (21.98)</td>
<td>100 (0)</td>
<td>90.00 (30.25)</td>
<td>83.33 (37.58)</td>
<td>94.72 (14.97)</td>
</tr>
<tr>
<td>20</td>
<td>90.00 (30.25)</td>
<td>96.67 (18.10)</td>
<td>90.00 (30.25)</td>
<td>100 (0)</td>
<td>85.00 (36.01)</td>
<td>78.33 (41.55)</td>
<td>90.00 (26.03)</td>
</tr>
</tbody>
</table>

Overall | 89.83 (22.56) | 94.75 (14.01) | 86.92 (26.55) | 88.58 (25.32) | 86.92 (27.24) | 89.25 (25.34) | 89.38 (23.50) |
Table 1 (cont.)

*Mean (SD) Student Engagement using 5-sec Time Sampling by 5-min Block*

<table>
<thead>
<tr>
<th>Student</th>
<th>Occasion 2</th>
<th>Block 1 0:05 – 5:00</th>
<th>Block 2 5:05 – 10:00</th>
<th>Block 3 10:05 – 15:00</th>
<th>Block 4 15:05 – 20:00</th>
<th>Block 5 20:05 – 25:00</th>
<th>Block 6 25:05 – 30:00</th>
<th>Overall 0:05 – 30:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>98.33 (12.91)</td>
<td>98.33 (12.91)</td>
<td>100 (0)</td>
<td>96.67 (18.10)</td>
<td>98.89 (7.32)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>90.00 (30.25)</td>
<td>98.33 (5.04)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>90.00 (30.25)</td>
<td>98.33 (5.04)</td>
<td>98.33 (5.04)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>88.33 (32.37)</td>
<td>95.00 (21.98)</td>
<td>90.00 (40.34)</td>
<td>83.33 (37.58)</td>
<td>93.33 (25.15)</td>
<td>91.67 (27.87)</td>
<td>88.61 (30.88)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>88.33 (32.37)</td>
<td>100 (0)</td>
<td>65.00 (48.10)</td>
<td>100 (0)</td>
<td>75.00 (43.67)</td>
<td>100 (0)</td>
<td>88.06 (20.69)</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>86.67 (34.28)</td>
<td>91.67 (27.87)</td>
<td>86.67 (34.28)</td>
<td>91.67 (27.87)</td>
<td>58.33 (49.72)</td>
<td>93.33 (25.15)</td>
<td>84.72 (33.20)</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>91.67 (27.87)</td>
<td>95.00 (21.98)</td>
<td>86.67 (34.28)</td>
<td>93.33 (25.15)</td>
<td>95.00 (21.98)</td>
<td>100 (0)</td>
<td>93.61 (21.88)</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>71.67 (45.44)</td>
<td>93.33 (25.15)</td>
<td>73.33 (44.59)</td>
<td>85.00 (36.01)</td>
<td>81.67 (39.02)</td>
<td>95.00 (21.98)</td>
<td>83.33 (35.37)</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>90.00 (30.25)</td>
<td>98.33 (12.91)</td>
<td>93.33 (25.15)</td>
<td>91.67 (27.87)</td>
<td>100 (0)</td>
<td>90.00 (30.25)</td>
<td>93.89 (21.07)</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>73.33 (44.59)</td>
<td>85.00 (36.01)</td>
<td>88.33 (32.37)</td>
<td>58.33 (49.72)</td>
<td>80.00 (40.34)</td>
<td>98.33 (12.91)</td>
<td>80.56 (35.99)</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>41.67 (49.72)</td>
<td>70.00 (46.21)</td>
<td>76.67 (42.65)</td>
<td>80.00 (40.34)</td>
<td>96.67 (18.10)</td>
<td>91.67 (27.87)</td>
<td>76.11 (37.48)</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>81.67 (39.02)</td>
<td>91.67 (27.87)</td>
<td>95.00 (21.98)</td>
<td>93.33 (25.15)</td>
<td>96.67 (18.10)</td>
<td>93.33 (25.15)</td>
<td>91.94 (26.21)</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>68.33 (46.91)</td>
<td>55.00 (50.17)</td>
<td>78.33 (41.55)</td>
<td>91.67 (27.87)</td>
<td>100 (0)</td>
<td>83.33 (37.58)</td>
<td>79.44 (34.01)</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>63.33 (48.60)</td>
<td>95.00 (21.98)</td>
<td>85.00 (36.01)</td>
<td>91.67 (27.87)</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>89.17 (22.41)</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>93.33 (25.15)</td>
<td>95.00 (21.98)</td>
<td>98.33 (12.91)</td>
<td>80.00 (40.34)</td>
<td>98.33 (12.91)</td>
<td>83.33 (37.58)</td>
<td>91.39 (25.15)</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>86.67 (34.28)</td>
<td>75.00 (43.67)</td>
<td>66.67 (47.54)</td>
<td>71.67 (45.44)</td>
<td>83.33 (37.58)</td>
<td>83.33 (37.58)</td>
<td>77.78 (41.02)</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>98.33 (12.91)</td>
<td>95.00 (21.98)</td>
<td>85.00 (36.01)</td>
<td>55.00 (50.17)</td>
<td>75.00 (43.67)</td>
<td>80.00 (40.34)</td>
<td>81.39 (34.18)</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>91.67 (27.87)</td>
<td>100 (0)</td>
<td>96.67 (18.10)</td>
<td>90.00 (30.25)</td>
<td>91.67 (27.87)</td>
<td>98.33 (12.91)</td>
<td>94.72 (19.50)</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>96.67 (18.10)</td>
<td>95.00 (21.98)</td>
<td>88.33 (32.37)</td>
<td>100 (0)</td>
<td>93.33 (25.15)</td>
<td>73.33 (44.59)</td>
<td>91.11 (23.70)</td>
</tr>
</tbody>
</table>

Overall 85.58 (27.49) 91.17 (21.34) 86.83 (29.16) 87.00 (27.03) 90.75 (21.07) 92.00 (22.15) 88.89 (24.71)
Table 2

*G Study Results for the Full Model (p x o) by Observation and Interval Length*

<table>
<thead>
<tr>
<th></th>
<th>10-min</th>
<th></th>
<th>15-min</th>
<th></th>
<th>20-min</th>
<th></th>
<th>30-min</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Var(^a)</td>
<td>% Var(^b)</td>
<td>Var(^a)</td>
<td>% Var(^b)</td>
<td>Var(^a)</td>
<td>% Var(^b)</td>
<td>Var(^a)</td>
<td>% Var(^b)</td>
</tr>
<tr>
<td>Person</td>
<td>78.28</td>
<td>75%</td>
<td>60.49</td>
<td>56%</td>
<td>51.83</td>
<td>52%</td>
<td>25.79</td>
<td>44%</td>
</tr>
<tr>
<td>Occasion</td>
<td>1.87</td>
<td>2%</td>
<td>0.61</td>
<td>1%</td>
<td>0.92</td>
<td>1%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Error</td>
<td>24.93</td>
<td>24%</td>
<td>47.54</td>
<td>44%</td>
<td>46.15</td>
<td>47%</td>
<td>33.2</td>
<td>56%</td>
</tr>
<tr>
<td>Person</td>
<td>107.37</td>
<td>70%</td>
<td>75.19</td>
<td>57%</td>
<td>68.36</td>
<td>59%</td>
<td>35.62</td>
<td>51%</td>
</tr>
<tr>
<td>Occasion</td>
<td>7.07</td>
<td>5%</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Error</td>
<td>39.18</td>
<td>26%</td>
<td>57.68</td>
<td>43%</td>
<td>46.68</td>
<td>41%</td>
<td>34.03</td>
<td>49%</td>
</tr>
<tr>
<td>Person</td>
<td>62.57</td>
<td>67%</td>
<td>39.67</td>
<td>39%</td>
<td>38.65</td>
<td>41%</td>
<td>20.76</td>
<td>36%</td>
</tr>
<tr>
<td>Occasion</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Error</td>
<td>30.65</td>
<td>33%</td>
<td>60.9</td>
<td>61%</td>
<td>54.83</td>
<td>59%</td>
<td>36.68</td>
<td>64%</td>
</tr>
<tr>
<td>N</td>
<td>84.88</td>
<td>62%</td>
<td>51.23</td>
<td>37%</td>
<td>58.16</td>
<td>48%</td>
<td>30.81</td>
<td>46%</td>
</tr>
<tr>
<td>Occasion</td>
<td>1.9</td>
<td>1%</td>
<td>0</td>
<td>0%</td>
<td>0.02</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Error</td>
<td>50.48</td>
<td>37%</td>
<td>85.62</td>
<td>63%</td>
<td>62.17</td>
<td>52%</td>
<td>36.06</td>
<td>54%</td>
</tr>
</tbody>
</table>

\(^a\) Var – variance calculated using Type II sum of squares

\(^b\) % Var – percentage of total variance
Table 3

*D Study Results for the Full Model (p x o) by Observation and Interval Length*

<table>
<thead>
<tr>
<th></th>
<th>Occasions</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10-sec</td>
<td>10-min</td>
<td>0.74</td>
<td>0.85</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>15-min</td>
<td>0.56</td>
<td>0.72</td>
<td>0.79</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>20-min</td>
<td>0.52</td>
<td>0.69</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>30-min</td>
<td>0.44</td>
<td>0.61</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>15-sec</td>
<td>10-min</td>
<td>0.70</td>
<td>0.82</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>15-min</td>
<td>0.57</td>
<td>0.72</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>20-min</td>
<td>0.59</td>
<td>0.75</td>
<td>0.81</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>30-min</td>
<td>0.51</td>
<td>0.68</td>
<td>0.76</td>
<td>0.81</td>
</tr>
<tr>
<td>20-sec</td>
<td>10-min</td>
<td>0.67</td>
<td>0.80</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>15-min</td>
<td>0.39</td>
<td>0.57</td>
<td>0.66</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>20-min</td>
<td>0.41</td>
<td>0.59</td>
<td>0.68</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>30-min</td>
<td>0.36</td>
<td>0.53</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>30-sec</td>
<td>10-min</td>
<td>0.62</td>
<td>0.76</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>15-min</td>
<td>0.37</td>
<td>0.54</td>
<td>0.64</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>20-min</td>
<td>0.48</td>
<td>0.65</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>30-min</td>
<td>0.46</td>
<td>0.63</td>
<td>0.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Figure 1. Percentage of variance attributable to person by observation and interval length