Peer Observation Methods to Estimate Classroom Behavior

John Hughes
Utah State University
Peer Observation Methods to Estimate Classroom Behavior

by

John Hughes

A plan-B paper submitted in partial fulfillment
of the requirements for the degree
of
Educational Specialist
in
Psychology
(School Psychology)

Approved:

UTAH STATE UNIVERSITY
Logan, Utah
2009
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>3</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>4</td>
</tr>
<tr>
<td>LITERATURE REVIEW</td>
<td>6</td>
</tr>
<tr>
<td>Description of Timed Sampling Observation Strategies</td>
<td>6</td>
</tr>
<tr>
<td>Validity and Accuracy in Time Sampling Observation Strategies</td>
<td>7</td>
</tr>
<tr>
<td>Systematic Time Sampling Observation Coding Schemes</td>
<td>10</td>
</tr>
<tr>
<td>Research on the Accuracy of Comparison Peer Techniques to Obtain Classroom Norms</td>
<td>14</td>
</tr>
<tr>
<td>Studies Collecting and Using Comparison Peer Data</td>
<td>16</td>
</tr>
<tr>
<td>METHODS</td>
<td>20</td>
</tr>
<tr>
<td>Systematic Review of Coding Scheme Manuals Recommending Comparison Peers</td>
<td>20</td>
</tr>
<tr>
<td>Inclusion and Exclusion Criteria of Manuals</td>
<td>20</td>
</tr>
<tr>
<td>Summary of Systematic Review of Coding Scheme Manuals Recommending Comparison Peers</td>
<td>23</td>
</tr>
<tr>
<td>Systematic Review of Studies Utilizing Coding Schemes and Comparison Peers</td>
<td>23</td>
</tr>
<tr>
<td>Inclusion and Exclusion Criteria of Studies</td>
<td>23</td>
</tr>
<tr>
<td>Areas of Evaluation and Coding Procedures for Selected Studies</td>
<td>23</td>
</tr>
<tr>
<td>RESULTS</td>
<td>24</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>37</td>
</tr>
<tr>
<td>Practical Implications for School Psychologists</td>
<td>41</td>
</tr>
<tr>
<td>Summary</td>
<td>45</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>47</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of Studies for each Observation Coding System</td>
<td>25</td>
</tr>
<tr>
<td>2. Number of Studies for each Purpose of Comparison Peer</td>
<td>26</td>
</tr>
<tr>
<td>3. Reported Peer Purpose for each Observation Coding System</td>
<td>27</td>
</tr>
<tr>
<td>4. Number of Studies Using each of the Different Types of Comparison Peer Strategies</td>
<td>28</td>
</tr>
<tr>
<td>5. Reported Peer Method for each Observation Coding System</td>
<td>29</td>
</tr>
<tr>
<td>6. Number of Studies and Length of Observation Session in Minutes</td>
<td>30</td>
</tr>
<tr>
<td>7. Comparison of Observation Code and Observation Length</td>
<td>30</td>
</tr>
<tr>
<td>8. Number of Studies Using Each Sampling Procedure</td>
<td>31</td>
</tr>
<tr>
<td>9. Number of Studies Using Each Length of Timing Interval in Seconds</td>
<td>32</td>
</tr>
<tr>
<td>10. Reported Interval Lengths for each of the Observation Coding Systems</td>
<td>33</td>
</tr>
<tr>
<td>11. Number of Studies and Types of Reliability Reported</td>
<td>33</td>
</tr>
<tr>
<td>12. Reported Reliability for each of the Observation Coding Systems</td>
<td>34</td>
</tr>
<tr>
<td>13. Number of Studies and Types of Validity Reported</td>
<td>35</td>
</tr>
<tr>
<td>14. Reported Type of Validity for each of the Observation Coding Systems</td>
<td>36</td>
</tr>
</tbody>
</table>
Introduction

School behavior observations are a data collection tool that allows systematic observation and coding of a student’s behavior. They are typically used to determine the degree that behavior affects academic learning. Behavior observations are a widely used and important part of a school evaluation. Observations are required by law for some special education evaluations (IDEIA, 2004) and are a common form of assessment tool used by school psychologists (Shapiro & Heick, 2004; Wilson & Reschly, 1996)

School psychologists use classroom observations to collect direct observational data on a student’s behavior in the learning environment. The data is often used to develop interventions that address problem behaviors. Direct observation data is also typically used as part of a comprehensive evaluation of a child who is experiencing behavioral or academic problems in the school setting. For instance, Eidle, Truscott, Meyers, and Boyd (1998), in a study of pre-referral interventions in elementary and secondary schools, found that around half of the student referrals had concerns of behavioral problems. Observation data is used in conjunction with other assessments to make important educational decisions such as a child’s eligibility for special education services. Observation data collected when a child is initially referred for behavior problems typically includes some sort of normative data to identify whether or not the child’s behavior is atypical of peers. Comparison peers, as normative data, have been used for a variety of purposes including as a benchmark to gauge if the target student’s behavior is different from the rest of the class (DuPaul & Hoff, 1998; Nelson & Bowles, 1975; Salvia & Hughes, 1990), to determine socially valid goals for effective classroom interventions (Walker & Hops, 1976), and to determine suitability of classroom management for maintaining appropriate classroom behavior (Jenson, Rhode, & Reavis, 1995; Walker & Hops, 1976).
School psychologists can choose from several different types of observational techniques to collect data on comparison peers. Time sampling techniques such as partial, whole or momentary are common practices to observe an individual’s behavior in the classroom. There are several student behavior coding schemes specifically developed for collecting and interpreting direct observational data in the classroom. Most of these coding systems also recommend collecting and using comparison peers to compare an individual child’s behavior with typical classroom behavior. Yet, knowledge on how accurate each of these techniques is in determining normative data is a critical factor when making important behavioral and educational decisions for a student. The Individuals with Disability Education Improvement Act of 2004 (IDEIA, 2004) requirements state that the assessment tools that are used in gathering information about a child should be “technically sound,” and used “for purposes for which the assessments or measures are valid and reliable” (P.L. 108-446 Part B sec 614 (b) (2and3)). Thus, it is critical for school psychologists to select an observation system that meets this requirement. Therefore, the first purpose of this paper is to summarize time sampling procedures that are commonly used in the classroom to collect data on student behavior. Second, research supporting the validity and reliability of time sampling techniques and systematic observational coding schemes will be discussed. Third, a review of manuals of four current structured behavior observation codes that recommend the collection of comparison peer data in the classroom setting will be presented. Fourth, the utility of comparison peer data will be further explored by reviewing studies that employed these observation systems. Finally, based on information from this review, practical implications for school psychologists in choosing accurate methods of data collection in behavior observations and future research directions for collecting classroom norms will be discussed.
Review of the Literature

*Description of Timed Sampling Observation Strategies*

When a student is referred to a school psychologist due to a classroom behavior problem, direct observation in the classroom is conducted as part of the data collected to make decisions about when and why the problem is occurring. Reasons for the problem occurring are then used to develop interventions or select services that would reduce problem behaviors and increase appropriate behaviors that support academic progress. Time sampling observational techniques are a common strategy used by school psychologists to measure and record student behavior in the classroom. A time sampling procedure involves observing whether a behavior occurs or does not occur during specified time periods to estimate behavior patterns or trends. During an observation session, the time period is separated into smaller intervals that are all equal in length. For example, a 10 minute observation session may be separated into 10 second intervals. The behavior is observed and recorded every 10 seconds to determine the percentage of intervals that a behavior was observed during the entire observation session.

Time sampling procedures are used because they can more readily estimate the duration and the number of occurrences of behaviors that are not discrete (no definite beginning or end). Three specific types of time sampling procedures include whole-interval recording, partial-interval recording, and momentary time sampling. Partial interval recording requires that an observer records a behavior as occurring if the behavior occurs at any time during each interval that is observed during an observation session. This type of procedure is used when behavior occurs for a brief amount of time making it difficult to catch and count. This procedure will not provide information about how many times a behavior occurs within an interval. Whole interval
recording requires that an observer records a behavior as occurring if the behavior occurred throughout the entire interval. Whole interval recording is useful when measuring continuous behaviors that occur for long periods of time to estimate both the duration of a behavior and the frequency of behavior occurrences during the observation session. During Momentary time sampling, an observer looks up to observe whether a behavior occurs or does not occur only at the end of an interval. The behavior that is observed at that moment in time is then recorded in a data sheet. Advantages of the momentary time sample process is that it does not require constant attention to a student's behavior all of the time and can be used to record frequent behaviors that the student engages in for a long period of time.

The purpose of all three time sampling procedures is to provide information about the occurrence of problem behavior which is used for identifying the problem, for analyzing the effects of environment on student behavior, and for developing interventions that adequately addresses the type of classroom problem that is occurring. Because reliability and validity determines the usefulness of this type of data to meet these purposes, the next section reviews the research on the psychometric properties of time sampling procedures.

*Validity and Accuracy in Time Sampling Observation Strategies*

In research studies, behavior observation's validity has generally been determined by estimating the accuracy of the coded behavior definition to encompass the "true state of the behavior as it actually occurs" when observing one person (Hintze & Matthews, 2004, p.259). Accuracy of true behavior occurrence is often determined by estimating the extent that the behavior observation code can correctly discriminate between children who are diagnosed with a behavior disorder and those demonstrating appropriate behavior (discriminate validity) or by
comparing the results of the observation code with an independent assessment that measures similar constructs (convergent validity).

It is virtually impossible for a single observer to accurately quantify an entire classes’ behavior using continuous observation for every peer at the same time to determine an accurate classroom norm. Thus, intermittent time sampling with a smaller sample of student behaviors to estimate classroom behavioral norms becomes necessary. Standard partial, whole, and momentary time-sampling procedures have been developed to conduct intermittent direct observations. These procedures can be used across different observers, time periods, and settings. Important to the validity of these measures is the accuracy or the degree that the measure reflects the behavior of interest. Thus, a valid sampled behavior would precisely measure the actual number of times a behavior occurs (frequency) the amount of time a behavior occurs (duration), or the amount of time it takes for a behavior to be initiated (latency). Studies have compared differing time sampling methods on individual behaviors to determine the most accurate recording procedure. Prior studies have found that momentary time-sampling, in which an observer codes the behavior occurring at the end of a time interval, is an accurate coding scheme for measuring low, intermediate, and high rate behaviors and that it reduces the error caused by method and by observer.

For example, Green, McCoy, Burns and Smith (1982), compared whole interval recording, partial interval recording, and momentary time-sampling techniques by observing a video tape of different rates of a target behavior (hair twisting). The rates of behavior were divided into occurrences at a rate of 25% (low), 50% (intermediate), or 75% (high) and replicated for each rate. The estimated occurrence of behavior based on results from the three time sampling procedures were compared to the three percentages of actual continuous measured
occurrence of hair twisting behavior. It was found that partial interval recording had a propensity to overestimate the actual number of behavior occurrence in all rates of behavior whereas whole interval recording method was more inclined to underestimate behavior at all rates. Alternatively, very little error occurred from the momentary time-sampling method. The authors found that most of the error came from the method of observation, instead of observer error. The results of this study indicated that momentary time-sampling yielded more accurate frequency results of behavior than partial or whole interval recording methods. Similarly, Murphy and Harrop (1994) found that a more accurate estimation of behavior occurrence was found when using results from a momentary time sampling as compared to using a partial interval time sampling procedure.

In a accuracy study of different momentary time-sampling intervals Keams, Edwards, and Tingstrom (1990) found that the shorter the interval the more accurate the coding. Kearns et al. used interval lengths of 30 seconds, 5 minutes, 10 minutes, and 20 minutes to code for behaviors that were simulated to occur at rates of 20%, 40%, 60%, and 80% of the time. Across all rates, the 30 second interval was more accurate than the longer intervals.

Although research suggests that a momentary sampling technique may be the most accurate frequency method to use in order to intermittently observe an individual’s classroom behavior, there is little empirical support to guide selection of the most accurate comparison peer data that reflects the actual classroom behavior environment. The purpose of comparison peer observations is to observe and record typical behaviors and to gain a true estimate of classroom behavior. If it is assumed that a child warranting intensive intervention is behaviorally “different” than peers, then identification of an at-risk child who differs from his or her peers seems to be dependent on the type of classroom environment. Consideration of the classroom
environment would enhance the number of true positive identifications of at-risk students. A child experiencing behavior difficulties due to poor classroom management is not likely to be different from peers.

*Systematic Time Sampling Observation Coding Schemes*

Due to the utility of behavioral observation for decision making, several observation coding scheme manuals are available to assess student behavior in the classroom. When evaluating the psychometric properties of these schemes, more research has been conducted on the reliability of systematic observation coding scheme measures than on the validity of these types of systems. Most often reliability is determined by the interobserver agreement - where two or more observations are collected concurrently and then compared for agreement. Generalizability theory, which attempts to determine reliability by factoring in all or most of the environmental issues that could be influencing outcomes, has also been used to determine reliability in direct observations. Factors that can influence observed behavior may include different settings or use of different methods, observers, or behavior dimensions (Hintze & Matthews, 2004).

Reed and Edelbrock (1983) conducted two studies to determine the validity and reliability of the Direct Observation Form (DOF) from the Child Behavior Checklist (CBCL). In the first study, the DOF was used to observe 25 boys age's six to eleven who had been referred for emotional and/or behavioral problems. Each boy was simultaneously observed six times for ten minutes at different times of the day over a two week period for a total of 60 minutes of observation for each target child. In the second study, six observers blind to the nature of the study and the clinical status of the children, observed 15 boys between the ages of six and eleven who were referred for problem behaviors with a sample of 15 non-referred boys from the same
classroom matched for grade, race and age. The teacher of each student in both studies also
completed a Teacher Report Form (TRF) for each of the observed students.

Reliability of the DOF measure was determined in two ways: interobserver agreement
and generalizability. Interobserver agreement was determined by comparing the results of the
first study. Agreement was high for both the total behavior problem and total on-task scores \( r = .91 \) and \( r = .83 \) respectively. Generalizability was determined by using a one-way intraclass
correlation and a two-way intraclass correlation for the results of the first study on both the total
behavior problem score and the on-task score. The one-way and two-way intraclass correlations
for the total behavior problem score were .86 and .93 respectively. The one-way and two-way
intraclass correlations for the on-task score were .71 and .83 respectively.

Convergent validity was determined by comparing the results of the observers' behavior
ratings on the DOF with the behavior ratings of the teacher on the TRF. A one-tailed test of
statistical significance comparing a total behavior problem score on the DOF and the total
behavior problem score on the TRF and the on-task score from the DOF with the school
performance score on the TRF resulted in significant similarities of \( p < .05 \) on all comparisons
except for two.

Discriminate validity was determined by contrasting the 15 referred boys on the second
study to their matched non-referred comparison peers on the total problem behavior score and
the on-task score from the DOF. On a one-tailed test of statistical significance the referred group
was significantly higher than the non-referred group on the total behavior problem scores (mean
of referred group = 41.5 +/- 3.7 versus mean of non-referred group = 27.0 +/- 2.8; paired t value
of 4.3 at a \( p < .01 \)). The non-referred group was significantly higher than the referred group on
the on-task score (mean of non-referred group = 53.4 +/- 1.4 versus mean of the referred group of 49.7 +/- 2.3; paired t value of 1.8 at a p < .05).

Hintze and Matthews (2004) used Generalizability Theory as a model for determining the reliability of systematic direct observations. Of the 14 participating students, three were receiving special education services. The observers, who were blind as to the classification of the students, used adapted definitions of on-task and off-task from the Behavioral Observation of Students in Schools (BOSS, Shapiro, 2004) to collect the behavioral data for the study. The behavioral data was collected twice a day during math and during language arts for 10 consecutive class days. Each student was observed using a momentary time sampling of 15 second intervals for 15 minutes during each instructional period. The resulting mean percentage of on-task behavior from the study was 80.2% with a standard deviation of 13.06. Interobserver agreement was calculated using 55 of the 245 available observations with each of the ten days of the study being represented. Agreement ranged from 88% to 92% with an average of 90%. Kappa indices were calculated to control for chance agreement. Kappa scores ranged from .31 to .93 with an overall average of .65 which demonstrated varying results from poor to good agreement.

The generalizability of this type of assessment was examined by estimating the variance components related to the students being observed (persons); the setting in which the observation took place; and the time of the observation. The greatest amount of variance was found in the category of persons with almost two thirds (62%) of the variability of the observation being explained by the variability of each observed child’s behavior as defined as “on-task” or “off-task.” The only other significant contribution to the variance was the interaction between persons and settings (around 13%). For instance some of the students were more on-task during math
than language arts and vice versa. Approximately 24% of the variance was considered a result of measurement error and could not be attributed to any particular component of the study.

Volpe, DiPerna, Hintze, and Shapiro (2005) reviewed the psychometrics properties of seven direct classroom observation coding systems including the Academic Engaged Time Code of the Systematic Screening for Behavior Disorders (AET-SSBD; Walker & Severson, 1990), the ADHD School Observation Code (ADHD-SOC; Gadow, Sprafkin, & Nolan, 1996), the Behavior Observation of Students in Schools (BOSS; Shapiro, 2004), the Classroom Observation Code (COC; Abikoff & Gittleman, 1985), the Direct Observation Form (DOF; Achenbach, 2001), the State-Event Classroom Observation System (SECOS; Saudargas, 1997), and the Student Observation System (SOS; Reynolds & Kamphaus, 2004). The study reported that all reviewed observation systems show some evidence that they accurately differentiate between children with behavioral difficulties and typically developing children. Only three of the seven behavior codes had evidence related to convergent validity. With the exception of one code, all had published interobserver agreement evidence of the coded behavior to support their use with school-age populations. Four of the seven had published evidence of treatment sensitivity, though much of this evidence is limited to single studies with samples that are small to moderate in size. As a result of these limitations in existing evidence, the authors recommended that much research is necessary to ensure that these codes are used for appropriate assessment purposes, behaviors, and target students. Until the completion of such studies, practitioners should select measures cautiously and use multiple methods in screening, diagnosis, and evaluation of treatment effectiveness.

In the review of observation systems, Volpe et al. (2005) emphasized the importance of using local norms. They found that the use of standard norms of typical classroom behavior is
difficult to achieve because of high levels of variability in important factors such as task demands, quality of instruction, or classroom behavior management. Comparisons of the reviewed seven scales revealed that five of the seven scales recommended some form of comparison peers as part of the observation system. Moreover, most systems recommended the collection of local normative data by assessing classroom peer behaviors under the same conditions as the target child.

Five of the seven codes examined in the Volpe et al. review suggested varying methods to observe comparison peers for determining local normative data. In the review on observation systems Volpe et al. (2005), did not report validity or reliability information on the varying comparison peer techniques to estimate classroom behavior norms. Although various types of comparison peers are recommended in these manuals and used in research studies, no study has compared the accuracy of various methods to the actual behavioral classroom norm. Thus, little is known about the reliability of decision making when different types of peer evaluations are used to determine local comparison norms.

Research on the Accuracy of Comparison Peer Techniques to Obtain Classroom Norms

Given that many classroom based coding systems recommend collecting and using comparison peer data, it appears to be an important part of the data collection process when making decisions about a student’s educational and behavioral program. Thus, studies on the psychometrics properties and utility of this type of data collection should be conducted. Bell and Barnett (1999) published one of the few research papers that have been written on the importance of comparison peers in the school setting for the purposes of goal setting, instructional design, progress monitoring, and eligibility determination. These authors posited that peer micronorms are "an effective technique for documenting discrepancies in child
behaviors and instructional strategies" (p.120) in the classroom. Peer micronorms, or comparison peer norms, are methods to collect useful data to determine these discrepancies in the classroom context when all students are participating in the same setting on the same instructional task. The authors also emphasized that technical adequacy criteria need to be fulfilled if peer micronorms are to be useful. The technical adequacy criteria include: 1. The sample size: which includes the amount of observations and/or the number of children sampled, 2. How the comparison peers are chosen and how the data are explained, 3. Normative statistics in one setting may not be adequate in another setting, 4. The significance of the differences between the target child and the comparison peers may be difficult to ascertain, 5. Appropriate comparison peers and well-functioning classrooms, and 6. The professional judgment of team members. These authors further state that comparison peer factors (where, when, and how to observe a certain number of comparison peers, the way peers are selected, etc.) may vary with the specific goal of observation.

Unfortunately, few empirical studies have explored the various factors that may influence observation accuracy and even fewer have been conducted in the last few decades. Hoier, McConnell, and Pallay (1987), for example, researched the accuracy of using two teacher nominated peers to obtain average classroom behavior. They used two teacher nominated “index” children with the criterion that they were “moderately successful, solid B students” (p. 8) and compared them to six (20% of the class) teacher nominated “average” students in the same classrooms. The results were reported as “reasonably good” (p.16) estimates of average classroom behavior; however, statistical similarity tests were not run on the results of the index children and the average peers because of the small number of classes studied.
Studies Collecting and Using Comparison Peer Data

Although few studies directly examined the psychometrics of various comparison peer data techniques, several studies used comparison peer data to investigate behavioral differences between at-risk or special education students experiencing behavioral difficulties or no risk students. A review of these studies may provide additional information about the utility of this type of data. For example, Skiba, McLeskey, Waldron, and Grizzle (1993) used comparison peers in a study on factors for referral of children to special education. These authors researched 24 male children in grades one through three in 24 different classrooms that were nominated by their teachers as "difficult-to-teach for either academic or behavioral reasons" (p. 84). They found that the referred students showed significantly more problem behaviors than their male peers. Behavior differences measured by the Code for Instructional Structure and Student Academic Response (CISSAR – Stanley and Greenwood 1980) included: less academic engaged time in reading (p<.05) and large group instruction (p<.05); also, more inappropriate behavior in reading (p<.05) and large group instruction (p<.01). Further, the at-risk students scored around a standard deviation lower than their peers on academic tests in reading and math (both: p<.001).

A study conducted by Deno (1980) indicated that the use of comparison peers is important to consider when doing a behavior observation because seemingly excessive behavior by a target student may be consistent with classroom norms. In this study, observations were conducted for students referred for services due to behavior problems. Observed problem behaviors of the referred students were frequent, but the problem behaviors were observed just as frequent in other students in the class. High percentages of problem behavior by many students suggests that the referred student may not be disruptive due to a disability but perhaps due to a lack of effective management of class wide behavior.
To determine the cause or function of the behavior problem, it is important to rule out poor classroom management as a major factor of behavior problems. This is accomplished by collecting observation data that indicates that the student is misbehaving in an appropriately managed classroom. There is evidence from a few studies indicating that most general education students are on task around 80% of the time in the classroom. For example, Forness and Esveldt (1975) utilized comparison peers in behavior observations to predict which kindergarten children would continue to have difficulty in the first grade. Using comparison peers, Forness and Esveldt (1975) showed that kindergarten children in four observed classrooms were on-task (total positive verbal and attending rates) 81.9% of the time during observations conducted from October through November and 84.9% during the months of February through March. Studies examining average classroom on-task behaviors have reported on-task ranges from 77% to 86% (Hintze & Matthews, 2004; Shin, Ramsey, Walker, Stieber, & O’Neill, 1987; Skiba et al., 1993). Moreover, the studies that used more than one classroom found that some classrooms had significant differences in on-task behavior between classes, illustrating the importance of developing local classroom norms.

Comparison peers are also used to gauge the effects of treatment that is designed to meet the normal limits of expected behavior in the classroom (Jenson, Rhode, & Reavis, 1995). For example, Walker and Hops (1976) used comparison peers to estimate typical classroom behavior that was set as a criterion measure for evaluation of treatment effectiveness. Three groups (seven female and seventeen male) of primary school students who were referred by their teachers as having poor classroom behavior, participated in the study. The collection of behavioral information was achieved through alternating momentary time sampling of the target child at the end of a six second time period, proceeding to a student peer, and then returning to the target
child. All peers in the classroom were observed by moving from one peer to another peer every subsequent six second peer observation period. Behavior observations indicated that the peers averaged from 60% to 80% in appropriate classroom behavior such as listening to instructions, following directions, attending to tasks, and completing assignments within the time allowed. Using this peer behavior data as the normal range of behavior during each observation session, the authors estimated an operant observation benchmark of expected appropriate behavior in the classroom across a variety of tasks. This operant measure could be used to track typical variability and trends in expected appropriate behavior over time as well as how participants' behavior variability corresponded with peer variability. Results revealed that the participants' behavior, like the comparison peers, was also variable; however, unlike the comparison peers their appropriate classroom behavior ranged between 20% and 40%. With the introduction of treatment, participants' appropriate behavior continued to vary but increased to a similar level as to his or her peers' benchmarks. Because observations were conducted across several tasks, when dips in participants' behavior corresponded with decreases in peer behavior, dips were attributed to classroom uncontrolled variability. Alternatively when decreases were lower than peer benchmarks during specific situations, explanations for the lack of generalization of treatment effects were considered. This systematic comparison of peer behavior served to determine realistic treatment goals and conditions under which the treatment goals were met.

When measures of behavior are used to make important educational decisions for a child, the utility of the measure is a significant concern. Reliability and validity evidence of the assessment process must be carefully considered in order to determine the level at which the assessment should be interpreted. As described above, researchers using comparison peers suggest that comparison peers have utility for deriving the nature of a child's problem behavior.
Thus, a systematic review of coding schemes that recommend comparison peers and of the literature using these systems is warranted to examine both psychometric properties and utility of comparison peer data. The purpose of the next section of this paper is to select and review manuals of current structured behavior observation codes that recommend the collection of comparison peer data in the classroom setting. Next, empirical support on the use of the selected observation systems with comparison peers will be explored by reviewing studies that employed these observation systems. Finally, based on information from this review, practical implications for school psychologists in choosing accurate methods of data collection in behavior observations and future research directions for collecting classroom norms will be discussed.
As part of the initial component of this review, current behavior observation coding systems that are used in the classroom were evaluated. To be included in this review, a behavior observation coding system must: 1. have been developed or updated in the last ten years; 2. been designed for use in classroom settings; 3. recommend or require the use of comparison peers; 4. have a published manual or administrative criterion; and 5. have been produced to observe general classroom behavior, not specific disabilities or disorders such as ADHD. Four formal behavior observation coding systems were reviewed including the Behavior Observation of Students in Schools (BOSS; Shapiro, 2004); the Direct Observation Form (DOF; Achenbach and Rescorla, 2001); the Revised Edition of the School Observation Coding System (REDSOCS; Jacobs et al., 2000); and the Student Observation System (SOS; Reynolds and Kamphaus, 2004).

Summary of Systematic Review of Coding Scheme Manuals Recommending Comparison Peers

The Behavior Observation of Students in Schools (BOSS; Shapiro, 2004) is a systematic direct observation coding system designed to assess academic skills behavior with the aid of a personal digital assistant (PDA). The time of the observation sessions with the BOSS can be adjusted to fit the need of the observer; however, the BOSS manual recommends observation lengths of at least ten minutes. During an observation session two student behaviors are collected: academic engagement and non-engagement. Engagement behaviors are coded using a momentary time sampling procedure at the end of a fifteen-second interval and non-engagement behaviors are coded using a partial interval recording procedure. Comparison peer observation is required every fifth interval to obtain classroom normative data. Although the specific number of
comparison peers is not indicated in the manual, Shapiro recommends that the observer use several children in the classroom in a rotating order (Volpe, DiPerna, Hintze, & Shapiro, 2005). No reliability or validity information is presented in the manual on comparison peer data.

The Direct Observation Form (DOF; Achenbach & Rescorla, 2001) is part of the Achenbach System of Empirically Based Assessment (ASEBA). The DOF is designed to assess on-task and problem behaviors in multiple settings including the classroom. Ten minute observation periods are utilized to write a narrative of the target child's behavior. The observation period is also broken into one minute intervals for a momentary time sampling of on-task behavior. At the end of the observation, the narrative and momentary time sampling results are used to score problem behaviors (a frequency count) and on-task behaviors (percentage). The DOF recommends that the same procedure be used to score two classroom "control" children directly before and after the target child's observation session to "identify ways in which the target child's behavior differs from that of peers observed under similar conditions" (Achenbach & Rescorla, 2001, p. 171). No reliability or validity information is presented in the manual on comparison peer data.

The Revised Edition of the School Observation Coding System (REDSOCS; Jacobs et al., 2000) is a systematic behavior observation code designed to evaluate disruptive behavior in the classroom setting. Whole and partial interval recording procedures are employed to calculate three behavior type percentages: appropriate versus inappropriate behavior, compliant versus noncompliant behavior, and on-task versus off-task behavior. Any number of children can be coded with the REDSOCS during an observation by rotating the observation focus from the target to a different classmate every sixty seconds until ten minutes of coding is complete for each target or peer. In the study supporting the psychometrics of the REDSOCS Jacobs et al.
(2000) used three randomly selected control children as comparison peers. No reliability or validity information is presented on comparison peer data for the REDSOCS.

The Student Observation System (SOS; Reynolds and Kamphaus, 2004) is part of the Behavior Assessment System for Children, Second Edition (BASC-II). The SOS was designed specifically for administration with children "whose emotional or behavioral problems are significant enough to impede academic progress" (Reynolds & Kamphaus, 2004, p. 49). The SOS utilizes a momentary time sampling procedure at the end of a thirty-second interval to indicate either adaptive or problem behaviors. At the end of the fifteen-minute observation period the frequency of specific behaviors are tallied. The SOS recommends collecting some "rough normative data" by completing the SOS on two or three same gender classroom peers of the target child. There is no specification of when the observations on the comparison peers are to be completed. No reliability or validity information is presented in the manual on comparison peer data.
Systematic Review of Studies Utilizing Coding Schemes and Comparison Peers

Inclusion and Exclusion Criteria of Studies

The second section of this review was to locate and analyze studies that used the selected observation systems. Articles for this review were selected if they met the inclusion criteria of using one of the four selected observation systems as a measure in their study and if the observation systems were used in a classroom setting. The principle source of literature used for this review originated through electronic databases such as ERIC, Psych Info, Psychology and Behavior Sciences Collection, and Google Scholar. Search terms included “behavior observation code;” “classroom observation;” “peer comparison;” “comparison peer,” and “classroom micro-norm.” Studies and reviews were searched back to 1980, with some background articles preceding 1980. Bibliographies of the selected studies were searched as well as articles that cited the included manuals for potentially relevant studies.

Areas of Evaluation and Coding Procedures for Selected Studies

Twenty-one studies that used one of the four selected behavior observation systems were found to meet the inclusion criteria. The author examined and coded each article. The number of target children along with the results of the behavior observations in the study for both the target and comparison children were recorded.

Data were collected from the articles to review information about how the behavior observation code was used in the study. First, the purpose or use of the comparison peer data was evaluated. The purpose of comparison peer data was coded as different if the stated purpose of the comparison peer observation was to differentiate the target child’s behavior from that of his classroom peers (DuPaul & Hoff, 1998; Salvia & Hughes, 1990), class management if used to determine if classroom management was suitable to maintain appropriate classroom behavior.
(Jenson, Rhode, & Reavis, 1995; Walker & Hops, 1976), or coded as intervention if used to help determine the effectiveness of classroom interventions relative to peers typical behavior (Walker & Hops, 1976). The peer purpose was also coded as a control if the purpose of the comparison peers was to establish a comparison group to study the validity of a treatment or that of the measure itself. If there were no comparison peers used in the study, the purpose was coded as no peers.

Second, the method of the comparison peers were recorded including comparison of only one peer, 2 to 4 peers, 5 or more peers, classwide (all students observed), a control group, or no peers used.

Information in regard to the observation time sampling procedures was also evaluated. For example, the sampling procedure of the observation was evaluated by coding the following types of sampling procedures: momentary time sampling, whole interval, partial interval, narrative, or none reported. In addition, the time of the observation interval used in each study was coded in seconds and the total time of the observation session was recorded in minutes.

Finally, reported psychometrics on the observation systems was reported. Reliability data was coded if reported and consisted of interobserver reliability, or generalizability. Validity data for the observation code was recorded as predictive, discriminate, convergent, divergent, or none according to the reporting of the study.

Results of the Systematic Review of Studies Utilizing Coding Schemes and Comparison Peers

The twenty-one studies were evaluated to determine how many of them utilized each of the four selected coding systems (see Figure 1). Eight articles (38% each) used the BOSS and the DOF, four (19%) used REDSOCS, and one article (5%) used the SOS. Of these studies, fourteen were group comparison studies and seven were single subject design studies. A range of 1 to 136
target students and a range of 0 to 227 peer students were included in the studies. A wide range of types of appropriate and inappropriate behaviors were coded across studies ranging from 2 categories of behavior to 96 specific behaviors. Eight studies examined significant differences between peers and target children and 100% of these studies reported significant differences between the two means of observed behavior(s) such that the measured target students’ behaviors were significantly more problematic (e.g., more disruptive) and/or less appropriate (e.g., lower on-task) than peers.

![Bar chart](image)

**Figure 1. Number of Studies for each Observation Coding System**

**Purpose of comparison peer data.** Each of the four behavior code manuals or authors of the codes stated a purpose for using comparison peers with their coding systems. The BOSS recommends obtaining local normative data through the use of comparison peers to understand the level that the behavior data found for the target student varies from the anticipated behavior of the student’s classroom peers (Shapiro, 2004). Although the DOF provides normative data on
287 children, it also recommends the use of randomly selected “control” peers to obtain a comparative classroom norm to “identify ways in which the target child’s behavior differs from that of peers observed under similar conditions” (p.171). The REDSOCS recommends using classroom control children to obtain a norm of behavior. The SOS recommends using selected comparison peers to obtain “rough” classroom norms (p.50).

Of the twenty-one studies, only two purposes for using comparison peers were found (see Figure 2). One study (4%) had two comparison peer groups: one group, from other classrooms, for a control group and the other group, from the same class as the target child, to see if there was a difference between the target children and their classroom peers. Thirteen studies, including the previous study, (59%) stated that the purpose of the comparison peer data was to serve as a control group comparison, and eight (37%) had no comparison peers. No study reported using the data to gauge the effectiveness of classroom interventions or to estimate overall classroom management.

![Figure 2. Number of Studies for each Purpose of Comparison Peer](image)

A comparison of the observation systems and the purposes of the comparison peers indicated that the BOSS was used in two studies and used comparison peer data as a control (See Figure 3). In addition, one of those studies had comparison peers in the same class to test
differences between the target children's behavior and that of their classroom peers. The BOSS was also used in six studies that did not use comparison peers. The DOF was used in seven studies. Six of these studies used comparison peer data as a control group and one study did not use comparison peers. The REDSOCS was used in four studies. Three of these studies used comparison peers as a control group and one study did not use comparison peers (single case study design). The SOS had a single study coded that used comparison peers as a control group.

![Figure 3: Reported Peer Purpose for each Observation Coding System](image)

*Comparison peer method.* Several different methods of selecting and observing comparison peers were recommended by each of the four codes. The BOSS recommends a sequential method of observing several comparison peers. To do this, the observer records the behavior of a classroom peer on every fifth interval going in an order that the observer chooses to code the behavior of each comparison peer and then rotates through the order again until the observation is complete (e.g., up and down rows). The DOF recommends completing the DOF on two randomly selected comparison peers. This is done by observing one peer before and another peer after the target child's observation session. The REDSOCS recommends using 3
randomly selected comparison peers in rotating sixty second intervals, starting with the first child for sixty-seconds, and then rotating to the next child, and so forth. The SOS recommends completing observations on two or three randomly selected same classroom peers either before or after the target child.

Of the twenty-one studies, two (9.5%) studies had a single comparison peer for each target child (See Figure 4). Two (9.5%) studies had two to four comparison peers, and one (4.8%) study had five comparison peers for each target child. There were eight (38.1%) studies that used a control group taken from outside the target children’s classroom and eight (38.1%) studies that did not use comparison peers. No study employed a class-wide comparison peer technique where all students in the class were observed.

Figure 4. Number of Studies Using each of the Different Types of Comparison Peer Strategies

A comparison of the observation systems and the comparison peer purposes indicated that six of the eight studies that utilized the BOSS did not use comparison peers (see Figure 5). Of the other two studies that used the BOSS, one used five same classroom peers per target child and the other used a control group. Four of the eight studies that used the DOF used a control
group, one used no comparison peers, two studies used one peer and one study used two-to-four comparison peers. The REDSOCS had one study that used two-to-four comparison peers, two studies that used control groups and one study that did not use comparison peers. The one study for the SOS used a control group.

![Bar chart showing the reported peer method for each Observation Coding System](image)

**Figure 5. Reported Peer Method for each Observation Coding System**

*Duration of an observation session.* Although most of the observation codes can be adjusted to fit the length of the observation desired, all codes recommended minimal time lengths of an observation session. The BOSS, DOF, and REDSOCS recommended an observation length of at least ten minutes; the SOS manual recommended an observation session length of fifteen minutes. A comparison of the observation systems and the comparison peer purposes indicated that of the twenty-one studies, nine (42.9%) used a ten minute observation length, seven (33.3%) used a fifteen minute observation, two (9.5%) used a thirty minute length, one (4.8%) study used a forty minute length, one (4.8%) had a forty-five minute observation and one (4.8%) did not report the length of the observation (See Figure 6).
Figure 6. Number of Studies and Length of Observation Session in Minutes

When comparing the lengths of the observation to each observation system, the BOSS had five studies that had a fifteen minute observation, and one study each that had a ten minute observation, and a forty-five minute observation (see Figure 7). There was also one study that did not report the length of the observation used. All of the DOF studies used a ten-minute observation length. The REDSOCS had two studies that used a thirty-minute observation length, and one study each that used a fifteen-minute and a forty-minute observation. The SOS had one study that used a fifteen minute observation length.

Figure 7. Comparison of Observation Code and Observation Length
Sampling procedure. Three of the observation codes (BOSS, DOF, and SOS) require the use of a momentary time sampling for at least part of their system. The BOSS also requires a partial interval recording method for recording off-task behaviors. The DOF requires producing a narrative of the target child's behavior for a sixty-second interval at the end of which the momentary time sampling procedure is used. The REDSOCS requires the use of partial and whole interval recording.

Of the twenty-one studies, thirteen (37.1%) reported using momentary time sampling, two (5.7%) mentioned using a whole interval procedure, eleven (31.4%) had partial interval, eight (22.9%) had a narrative procedure as part of the study, and one (2.9%) study did not mention what time sampling procedure was used (see Figure 8). Eight studies had both narrative and momentary time sampling, four had momentary time sampling and partial interval sampling and two studies had partial and whole interval sampling.

![Figure 8. Number of Studies Using Each Sampling Procedure](image)

Interval time using a time sampling procedure. All four observation codes had different observation interval lengths for recording behaviors using a time sampling procedure. Specifically, the REDSOCS recommends using a ten second interval, the BOSS recommends
using a fifteen second interval, the SOS recommends using a thirty second interval, and the DOF recommends using a 60 second interval to record behaviors. Of the twenty-one studies, two (9.5%) did not report the length of intervals used, four (19%) used a ten second interval, six (28.6%) used a fifteen second interval, one (4.8%) study used a thirty second interval, and eight (38.1%) used a sixty second interval (see Figure 9).

**Figure 9. Number of Studies Using Each Length of Timing Interval in Seconds**

Figure 10 displays the percentage and frequency of each interval time reported in the reviewed studies. As shown in Figure 10, each of the studies that reported the length of the intervals followed the interval recommended by the coding system. There were two studies that used the BOSS that did not report the length of intervals used; however, they did not mention any deviation for interval length from the regular administration.
Figure 10. Reported Interval Lengths for each of the Observation Coding Systems

Reliability. Only two types of reliability data were reported in the reviewed studies (see Figure 11). One-hundred percent of the studies reported reliability data for the observation codes used. All of the studies reported interobserver reliability numbers with correlation means ranging from .71 to .95, while one study that used the DOF (Reed and Edelbrock, 1983) also reported generalizability numbers ranging from .86 to .93. (see Figure 12).
Validity. Nine of the twenty-one studies looked at the validity of the observation system and four of those nine looked at more than one type of validity (see figure 13). Twenty-six validity types were coded with twelve (46.2%) of those twenty-six reporting no validity. Most of the studies (8-30.8%) that looked at validity were looking for discriminate validity. There was one (3.8%) study that looked for predictive validity, four (15.4%) studies that looked for convergent validity and one (3.8%) study that looked at divergent validity.

All of the studies that looked at discriminate validity (n=8) found significant differences between referred problem behavior target children (such as those diagnosed with ADHD) and "normal" comparison peers (at least p<0.05). In addition to discriminative validity, four (15.4%) studies evaluated convergent validity. Three studies used the REDSOCS. One found significant negative correlations between the percentage of appropriate behavior on the REDSOCS and the conduct problem, hyperactivity, and hyperactivity index subscales of the Conners' Teacher Rating Scale (Filcheck, Berry, & McNeil, 2004). Another found significant correlations between the REDSOCS and the Sutter-Eyberg Student Behavior Inventory and the Revised Conners'
Teacher Rating Scale (Jacobs et al., 2000). The last one found significant correlations between
the adjusted REDSOCS categories and the Adjustment Scales for Preschool Intervention
(Bulotsky-Shearer & Fantuzzo, 2004). Reed & Edelbrock (1983) found significant correlations
between the DOF and the Teacher Report Form. One (3.8%) of the twenty-one studies
(Filcheck, Berry, & McNeil, 2004) also evaluated divergent validity and found no correlations
between the REDSOCS with the Conners' Teacher Rating Scale's anxious-passive subscale.
Finally, only one study (Lett & Kamphaus, 1997) evaluated predictive validity of the SOS's
ability to correctly predict children diagnosed with ADHD. The study reported that the SOS does
not significantly contribute to the predictive validity of the Teacher Rating Scale.

Figure 13. Number of Studies and Types of Validity Reported

Of the observation systems the BOSS had six studies that did not obtain validity for the
system itself and two studies that validated the system through discriminate validity (see figure
14). The one SOS study used both discriminate and predictive validity. The DOF had five studies
that did not validate the observation system itself, three studies that validated the system through
discriminate validity, and one of the discriminate validity studies also validated the DOF through
convergent validity. The REDSOCS had one study that did not look at validity for the
observation system, one study that looked at discriminate, convergent, and divergent validities, one study that looked at convergent and discriminate validity, and one study that looked at discriminate validity.

![Figure 14. Reported Type of Validity for each of the Observation Coding Systems](image-url)
Discussion of Systematic Reviews

One purpose of this paper was to select and review manuals of current structured behavior observation codes that recommend the collection of comparison peer data in the classroom setting. Four behavior coding systems were found to meet the inclusion criteria of this study. Although all four observation systems suggest obtaining classroom behavior norms by observing comparison peers, the method of observation varied. Two of the observation systems (BOSS & REDSOCS) recommend collecting data on comparison peers during the observation of the target child. Both the DOF and the SOS recommend collecting data on comparison peers either before or after the observation of the target child. The recommended number of comparison peers also varied per system. The BOSS recommends “several” peers, the REDSOCS recommends three or more, the DOF recommends two, and the SOS two or three.

Information derived from a review of the manuals revealed several strengths. For example, three of the four systems (BOSS, DOF & SOS) utilized momentary time sampling for at least part of their comparison peer observation, which has been shown to accurately estimate behavior (Green, McCoy, Burns, & Smith, 1982; Kearns, Edwards, & Tingstrom, 1990; Murphy & Harrop, 1994). Two of the systems (BOSS & REDSOCS) are able to observe comparison peers during the observation of the target child. By simultaneously observing the target and comparison peers it allows the observation systems to control for variations in classroom behavior (Walker & Hops, 1976).

Several noteworthy limitations were also apparent. First, the systems were not always clear on the procedures that they recommended to collect comparison peer data. For instance the number of comparison peers and time of their observation were sometimes unclear. The BOSS only stated that “several” comparison peers should be used. Also, although three comparison...
peers were used to validate the REDSOCS, the administration criteria did not recommend a certain number of comparison peers. As for the time of the observation, the SOS did not clarify when comparison peers should be observed in relation to the target child’s observation. Second, no information was provided in the manuals about the validity and reliability of the recommended peer evaluations. This lack of psychometric data makes it difficult for school psychologists to have confidence in the ability of the varying comparison peer techniques to accurately determine classroom behavior norms.

Another purpose of this paper was to systematically review studies to further clarify the observation methods, procedures, purpose, and available psychometric properties of the four current systematic observation codes used to assess peer classroom behavior. One positive finding was that all twenty-one studies that were reviewed reported reliability findings for the behavior coding system that was used. All of the studies reported inter-observer agreement numbers while one of the studies (Reed & Edelbrock, 1983) also reported generalizability numbers for reliability. All studies reported correlation means which ranged from .71 to .95; however, they did not separate target and peer reliability data.

Reported validity on the four observation coding systems was also reviewed in the studies. Nine of the twenty-one studies looked at the validity of the observation system and four of those nine looked at more than one type of validity. Most of the studies that looked at validity were evaluating discriminate validity. All of the studies that looked at discriminate validity (n=8) found significant differences between referred problem behavior target children (such as those diagnosed with ADHD) and “normal” control groups (at least p<0.05). Five of those eight studies used peers in the same class as the target children. Further, three of those five studies reported using the comparison peer technique recommended by the observation manual.
In the VileJunod et al. (2006) study, the BOSS system was used to observe both a control group outside of the classroom and classroom comparison peers to obtain a behavioral difference between peers and the target child. The two remaining studies that used comparison peers in the same classroom (Skansgaurd & Burns, 1998; Reed & Edelbrock, 1983) both used the DOF system but used a single peer instead of the two that are recommended in the DOF manual. Reed and Edelbrock (1983) used two observers to complete the DOF on the target child and the comparison peer simultaneously and Skansgaurd and Burns (1998) observed the target and comparison peer one right after the other as opposed to one before and one after the target child as the manual recommends. The other three studies that researched discriminate validity used children from different classes or schools. Thus, although these studies evaluated validity for the BOSS, REDSOCS and SOS systems, this evaluation was conducted using procedures that differed from those recommended by the systems.

In addition to discriminative validity, four (15.4%) studies evaluated convergent validity. Three studies used the REDSOCS. One found significant negative correlations between the percentage of appropriate behavior on the REDSOCS and the conduct problem, hyperactivity, and hyperactivity index subscales of the Conners' Teacher Rating Scale (Filcheck, Berry, & McNeil, 2004). Another found significant correlations between the REDSOCS and the Sutter-Eyberg Student Behavior Inventory and the Revised Conners' Teacher Rating Scale (Jacobs et al., 2000). The last one found significant correlations between the adjusted REDSOCS categories and the Adjustment Scales for Preschool Intervention (Bulotsky-Shearer & Fantuzzo, 2004). Reed & Edelbrock (1983) found significant correlations between the DOF and the Teacher Report Form. One of the twenty-one studies (Filcheck, Berry, & McNeil, 2004) also evaluated
divergent validity and found no correlations between the REDSOCS with the Conners' Teacher Rating Scale's anxious-passive subscale. Finally, only one study (Lett & Kamphaus, 1997) evaluated predictive validity of the SOS's ability to correctly predict children diagnosed with ADHD; however, the study reported that the SOS does not significantly contribute to the predictive validity of the Teacher Rating Scale.

It is important to note that like the manuals of the coding systems, none of the studies reported validity specifically for the comparison peer technique used, and its accuracy to obtain a classroom or local norm. However, the studies were typically not trying to validate the coding system and used acceptable control group techniques that controlled for variables such as age, grade, gender, race, and socioeconomic status between the target children and the control group children. Only nine of the twenty-one studies found to fit the inclusion criteria for this review, looked at validity for the four behavior coding systems. Of those nine several reported using adapted versions of the behavior coding system and only three used the comparison peer techniques recommended by the manuals. Clearly there is a need to further validate these behavior observation systems, specifically in regards to the comparison peer techniques recommended.

There were no studies, or manuals found on these specific coding systems that validated their technique of gathering local or classroom norms from comparison peers; however, there were a few studies and reviews found in the background search that attempted to validate comparison peer techniques (Bell & Barnett, 1999; Hoier, McConnell, & Pallay, 1987). The comparisons of their findings with the recommended techniques in the coding manuals are evaluated in the implications for school psychologists section. Although comparison peers are purported to be important data to make conclusions in regard to a target students' problem
behavior being exceptional and not due to poor classroom management, this research paper has found that the existing evidence for current comparison peer techniques used by the four manuals that met inclusion criteria is limited at best. In order to help school psychologists exercise best practice techniques in creating classroom norms of behavior, evidence supporting pieces of the comparison peer techniques will be discussed and future research directions explored.

Practical Implications for School Psychologists

The final purpose of this review was to discuss practical implications for school psychologists in choosing accurate methods of data collection in behavior observations and future research directions for collecting classroom norms. One initial question for school psychologists when conducting comparison peer observations naturally will be, "how do I choose appropriate comparison peers?" This question is answered in two parts: the first is the evidence for choosing comparison peers and the second is how many comparison peers should be used.

All four of the reviewed coding systems recommend using varying numbers of randomly selected peers usually of the same gender as the target child; however, random selection of comparison peers to obtain a classroom norm has no published support. Hsu (1989) recommended when using an unconstrained randomization sample that the sample should be at least twenty subjects. This sample size would actually approach the size of the population of the class. A more realistic suggestion for classroom random sampling comes from Tindal (1988). He suggested "a minimum of 15-20% of the student population needs to be sampled to produce a stable index of normative performance" (p. 119). According to the U.S. Department of Education, (2002) an average public class consists of 21.1 to 23.6 students. That would equate to around four or five comparison peers per average classroom. The SOS recommends using two or
three randomly selected same gender comparison peers, and the DOF recommends using two randomly selected comparison peers to estimate a classroom norm. These sample sizes do not equate to fifteen to twenty percent of an average class. The REDSOCS recommended three comparison peers in (Jacobs et al., 2000); however, one of the strengths of the REDSOCS is that it facilitates the use of as many comparison peers as the observer desires during the observation of the target child. The BOSS also can accommodate unlimited comparison peers during the target child’s observation.

Although the BOSS only recommends observing “several” comparison peers, it’s approach to viewing and coding comparison peers in the class is the same as the whole-room technique recommended by the Manual for Coding Discrete Behaviors in the School Setting (MCDB; Patterson, Cobb, & Ray, 1972). This technique would increase the sample size to the total number of same gender peers in the classroom. Face validity would indicate that this technique would be the most accurate to obtain classroom normative data; however, there was no published validity data found for this approach.

Although teacher nomination of “average” comparison peers was not one of the techniques recommended in the four coding systems reviewed, it is recommended in some literature and used in national coding systems (e.g., Classroom Observation Code; Abikoff & Gittelman, 1985). Teacher nomination of a “normal” child in the classroom has received some evidence for reliability (Bell & Barnett, 1999; Hoier, McConnell, & Pallay, 1987; Strain, Lambert, Kerr, Stagg, & Lenkner, 1983), but the evidence is limited (Hoier, McConnell, & Pallay, 1987) and is usually focused on the teacher nominating high and low academically and socially competent children in the classroom and not “average” behaving children (Strain et al., 1983).
Another issue is the selection of a time sampling procedure. Important to the validity of these measures is the accuracy or the degree that the measure reflects the behavior of interest. Thus, an accurate sampled behavior would precisely measure the actual number of times a behavior occurs (frequency), the amount of time a behavior occurs (duration), or the amount of time it takes for a behavior to be initiated (latency). Studies that have compared these methods on individual behaviors to determine the most accurate intermittent recording procedure have found that momentary time sampling is the most accurate method versus partial or whole interval methods (Green, McCoy, Burns, & Smith, 1982; Kearns, Edwards, & Tingstrom, 1990; Murphy & Harrop, 1994). Three of the four coding systems, including the DOF, SOS, and BOSS, recommended using momentary time sampling to code behaviors in the class. The BOSS also recommends using a partial interval method for coding off-task behavior and the DOF has a narrative component. Only the REDSOCS does not recommend momentary time sampling; instead, it recommends the use of partial and whole interval methods.

Kearns, Edwards, and Tingstrom (1990) found that the accuracy of momentary time sampling increases with smaller intervals. For the three systems that use momentary time sampling the DOF has an interval length of 60 seconds, the SOS has an interval of 30 seconds and the BOSS has an interval of 15 seconds; however, comparison peer data is only taken every fifth interval on the BOSS which produces an interval of 75 seconds for comparison data. Alternating small intervals between the target child and comparison peers would decrease the interval length and hopefully increase the accuracy of the normative data.

One of the main purposes of using comparison peers is to control for variations in classroom behavior and behavioral expectations (Walker & Hops, 1976). In order to control for these variations and get an accurate classroom norm for the time in which the target student was
observed, an observer must collect simultaneous comparison peer data. The BOSS and the REDSOCS instruct the observer to do this; however, the DOF and the SOS cannot be completed on more than one child without increasing the number of observers. Completing Behavior observations of comparison peers before or after the observation of the target child does not effectively control for time variables in expected behaviors. For example a target child might be observed during independent math work, while a comparison peer observed after the target child may be observed during small group reading instruction. Both activities would typically have different behavioral expectations and interest levels for the students.

In relation to the four observation codes, lack of research supporting comparison peer recommendations makes it difficult to select an observation coding system for conducting comparison peer data collection. For example, although several coding systems recommended same gender comparison peers, gender differences in behavior were not researched. Gender and other possible variables (such as race, and socioeconomic status) that could result in behavior differences between samples were not addressed in this review to keep the focus on evidence validating varying comparison peer techniques that would accurately determine a classroom norm, not to determine a valid comparison group to the target child. In particular, clear conclusions were also difficult to ascertain for the REDSOCS because some of the studies were using adapted versions of the coding system. All reviewed studies using REDSOCS except one reported using a modified version of the REDSOCS.

In terms of what methods and procedures were most frequently used in the studies, evidence from this review shows that observations of comparison peers primarily utilized one to four comparison peers during a 10 to 15 minute observation session using a momentary time sampling procedure with a 60 second interval. There are indications that investigations should
focus on methods that incorporate a comparison peer selection that includes a whole room approach, more than four randomly selected comparison peers, or several teacher nominated “average” peers. To control for time variables, simultaneous observation of the comparison peers and the target child are also recommended, and studies indicate that a momentary time sampling method with short intervals may produce the most accurate outcomes. The comparison peer technique recommended by the BOSS fits many of the supported research criteria and potential purposes of comparison peer data. Specifically, this system simultaneously collects target and comparison peer data, and has the ability to observe more than four randomly selected peers or a whole room technique to potentially estimate classwide behavior (Shapiro, 2004). It also uses the less biased momentary time sampling method with a fifteen second interval (seventy-five seconds for comparison peers) to determine academic engaged behavior; however, it uses a partial interval method to record off-task behavior, which the manual admits is probable to overestimate occurrence (ibid). A primary disadvantage of the BOSS was that most of the studies that used the BOSS in this review, did not obtain validity data for it.

Summary

In summary, comparison peer techniques have great potential to aid school psychologists and other school professionals in making more careful and valid decisions about the nature of a child’s problem behavior by gauging typical behavior across numerous educational tasks and situations but may not be captured with standard normative data such as national norms (Deno, 1980; Jenson, Rhode, & Reavis, 1995; Walker & Hops, 1976). Although there are many recommended techniques to choose and collect local normative data from comparison peers in current behavior observation codes, there is no evidence from coding manuals or studies that the recommended comparison peer techniques accurately determine classroom norms. As a result of
these limitations, much research is necessary to provide reliability and validity evidence supporting comparison peers. Specifically, comparison peers use to document discrepancies between target child behaviors and peers within the context of classroom routines to determine classwide management, intervention effectiveness, or eligibility determination. Meanwhile, it is clear that comparison peer data when used should be interpreted with great caution. Furthermore, it is important that school psychologists take additional steps besides those recommended in the manuals to maximize the validity and the utility of observational assessments. For example, multiple observations with children doing similar tasks achieve a more reliable estimate of a student’s behavior (Hintze, 2005). In addition, multiple observations are likely to decrease student reactivity to the observer. Moreover, this data should be used in conjunction with additional data that provides information about classroom management and norms (e.g. school-wide discipline referrals). The conclusions of this review will hopefully be a springboard for investigations that will address the lack of research on comparison peer methods to obtain classroom or local micronorms.
References


Individuals with Disabilities Education Improvement Act of 2004 (PL 108-446) 20 USC 1400.


*Indicates articles used in research review*