1-2005

Observations and ratings of preschool children's social behavior: Issues of representativeness and validity

Brian McKevitt
University of Nebraska at Omaha, bmckevitt@unomaha.edu

Stephen N. Elliott
Vanderbilt University

Observations and ratings of preschool children's social behavior: Issues of representativeness and validity

By:

Brian C. McKevitt
Heartland Area Education Agency 11, Johnston, IA

and

Stephen N. Elliott
Peabody College at Vanderbilt University

Data were gathered from videotaped recordings of two preschool children engaged in unstructured free play over 12 days each. Observers coded behavior from the videotapes and completed a behavior rating scale for each child after every two observation sessions. Teachers also completed two behavior rating scales per child. Results indicated that at least three 30-min observation sessions were required to reliably represent a child's overall behavior. Moderate correlations were obtained when observations were compared with teachers' and observers' own ratings, indicating the behavior rating scale did an adequate job of reflecting actual observed behavior. The implications of these results for researchers and practitioners are discussed.

Best practices in assessment call for a multimethod, multisource approach; that is, using several methods of assessment that vary by informant, structure, and format to obtain a comprehensive view of children's emotional and behavioral functioning (McConaughy & Ritter, 1995). Information from multiple sources and multiple methods is then integrated and used to make important decisions about a child's functioning and educational program. Observations and rating scales are two examples of assessment methods commonly used in such an approach. Rarely is a question asked about how much information from these measures is needed to ensure that a representative "picture" of behavior has been obtained. Additionally, it is often assumed that different assessment methods addressing similar behaviors yield results that are in high agreement

Data are limited regarding how much observational data one should collect to gain a representative sample of behavior. Barton and Ascione (1984) reported that in general, the more data that can be gathered, the better. Too few data could threaten the validity and reliability of the observations. Shapiro (1996) wrote that a "best guess" recommendation for how much observation data should be collected is to observe for at least one full period in which problems exist, and then portions of that period on one or two other days. Doll and Elliott (1994) addressed the issue of how much observational data are enough by using a correlational research design to examine the number of classroom observations required to gain an accurate and representative sample of a preschool child's social behavior. Using 24 children observed over six weeks, the authors compared early observation sessions to later sessions using correlations and kappa coefficients, and compared the results of a complete set of nine observation sessions to those of the first session, the first two sessions, the first three sessions, and so on. Results from these comparisons indicated that neither two nor three observation sessions were

sufficient to describe a consistent pattern of social behavior. After five observations, six of eight behaviors correlated highly (*r* = .80) with the total observation record. From these data, the authors concluded that at least five 20-min observation sessions across several weeks would adequately represent a preschooler's social behavior. In addition, they found that the type of behavior accounted for the variation in predictability of behaviors. Some behaviors, such as directed play or physical aggression, were much more consistent in their occurrence than others.

Another source of variance stems from different informants. Achenbach, McConaughy, and Howell (1987) performed a meta-analysis to study agreement among informants' ratings of children's behavior. In a review of 119 studies, they found that the mean *rs* between all types of informants were statistically significant. Similar informants (i.e., pairs of teachers, pairs of mental health workers) had the highest correlations (mean *rs* = .64 and .54, respectively). Informants with different roles (i.e., teacher–parent pairs) had lower correlations, but still significant, with the highest occurring between teacher and observer pairs (mean *r* = .42). Mean agreement between pairs of observers was *r* = .57.

Though different informants using the same assessment method can have significant levels of agreement, it is another issue to conclude different assessment methods also share consistency. Merrell (1993b) found that correlations between the Child Behavior Checklist-Direct Observation Form (CBC-DOF; Achenbach & Edelbrock, 1986) and the School Social Behavior Scale (SSBS; Merrell, 1993a) were weak to moderate between teachers and observers for problem behavior scores (*r* = -.06 – -.39) and moderate for on-task ratings (*r* = .26 – -.52). Likewise, Elliott, Gresham, Freeman, and McCloskey (1988) found that teachers' and observers' ratings on the Social Skills Rating System-Teacher Version (SSRS-T; Gresham & Elliott, 1990) and observers' observations correlated moderately for certain categories of observed behavior.

Research which used observation systems designed to be used with a specific rating scale also has found moderate correlations between the two methods. Robertson (1993) and Racine (1994) both researched the relationship between observations and ratings using an observation system designed by Robertson to be used with the SSRS-T. Robertson compared teachers' ratings on the SSRS-T to observers' observations on the observation system, teachers' ratings to observers' ratings, and observers' ratings to other observers' ratings. Ratings and observations of children were completed via videotaped vignettes. Robertson found moderate correlations between teachers' and observers' ratings (mean *r* = .58), yet low to moderate correlations (mean *r* = .38) between teachers' ratings and observers' observations.

In a similar study using videotaped vignettes of child behavior, Racine (1994) compared observations to rating scales, but only used observers' observations and ratings. Racine found that observers' observation records correlated moderately with the same observers' ratings. Mean correlations on the two videotaped vignettes were *r* = .59 and *r* = .56.

A limitation of the studies by Racine (1994) and Robertson (1993) is that their correlations were based on one observation session and one behavior rating for each video. The question still exists whether a longer observation record would yield higher correlations between observers' ratings and observations. Additionally, the observation system used by Robertson and Racine includes only a subset of behaviors rated on the SSRS-T. This may contribute to the moderate correlations found in their studies because the rating scale addresses more behaviors than the observations. Finally, they may have chosen a sample of time to videotape that was not representative of children's behavior.

As a result of the concerns with the aforementioned studies, several practical questions were stimulated and addressed: Because the observation system includes only a subset of behaviors rated on the scale, will correlations increase if only those behaviors that are rated and observed are compared? Will the correlations between total observations and ratings increase if there were more data from each method? If so, is it possible to say that after a certain number of observations, the rating scale adequately represents observed behavior? How many observations are required to gain such an adequate representation?

Method

Participants

Two preschool children enrolled in an early childhood program participated in this study. One male child with a disability (Down's syndrome; Child 1) and 1 female child without a disability (Child 2) were included in this study to represent children who might be observed in many early childhood programs.

Written parental consent for each child and participation consent from the five teachers working in the classroom were obtained. All teachers had been working with the children in the classroom for three months at the time the videotaping began. Three experienced teachers volunteered to complete behavior ratings for children being videotaped. Teacher 1 completed ratings for both children, Teacher 2 completed ratings only for Child 1, and Teacher 3 completed ratings only for Child 2.

Two educational psychology graduate students performed the observations for this study. These students were recruited by direct verbal contact, and their participation was voluntary. Both observers had prior training in behavioral observation and received additional training with the observation system used in this study. Written consent was obtained from each observer after they were provided a summary of the nature and the purpose of this investigation.

Materials

The materials used in this study included the preschool form of the SSRS-T (Gresham & Elliott, 1990), an observation system designed to target four behavior categories identified on the SSRS-T, and videotapes of children engaged in spontaneous play.

The SSRS is a multirater assessment instrument used by teachers, parents, and children to examine children's social behaviors. Teachers and other school professionals use the SSRS-T to rate the frequency of children's social behaviors and the importance they attribute to those behaviors. This information is then used to evaluate children's social competence and classroom functioning. The Teacher Version of the SSRS for the preschool level includes two major scales: Social Skills and Problem Behaviors. Social Skills subscales include Cooperation, Assertion, and Self- Control. Problem Behavior subscales include Externalizing, Internalizing, and Hyperactivity. Frequency and perceived importance of social skill and problem behaviors are rated on a 3-point Likert-type scale. The reliability and validity evidence for the SSRS is substantial (see the SSRS Technical Manual; Gresham & Elliott, 1990). A recent comparative review of published social skills instruments concluded it has among the best overall psychometric characteristics of the instruments reviewed (Demaray et al., 1995).

Observers used the observation system designed by Robertson (1993) for use with the SSRS-T. Information regarding the operational definitions of observed behaviors and their correspondence with those rated on the SSRS-T may be provided by the author upon request. Inclusion of items in the observation system was based on the following criteria: (a) top factor loading items from each subscale on the SSRS-T (Cooperation, Assertion, Self-Control, and Problem Behaviors), (b) singular behaviors, (c) discrete observable behaviors, and (d) observable classroom behaviors that occur with reasonable frequency. The observation system uses an event recording format in which observers record frequency of behaviors (occurrence) over the duration of the observation period. Observers also may document opportunities for certain behaviors to occur, such as those that are contingent upon the occurrence of other events (e.g., "Follows teacher directions" is contingent upon a teacher giving a directive; "Inviting others to join" is contingent upon play that lends itself to inviting others to join).

Procedure

Videotaping. Children were videotaped by the researcher during spontaneous play periods in their classroom over two months. Child 1 was videotaped first over 12 consecutive days of attendance, followed by Child 2 for 12 consecutive days. Videotaping did not occur when the children arrived late for free play, were absent, or when school was not in session. Each videotaping session lasted an average of 32 min. Each child's play was recorded during the entire session; recording paused only if the child moved to another part of the room and the researcher had to reposition himself and the camera. This real-time recording of behavior reduced the chance for systematic sampling error that could affect the validity of the observations (Foster & Cone, 1995).

Observation training. Observers were trained by the researcher after videotaping was completed. Observer training consisted of two stages conducted over two 2-hr sessions. First, observers learned and discussed definitions of target behaviors in the observation system and applied the definitions to written descriptions of child behavior. Second, observers viewed two videotapes of preschool children involved in free play and coded their social behaviors using the observation system. The videotapes the observers viewed during the training were the same videotapes used by Racine (1994) and Robertson (1993) in their studies. Interobserver reliability was calculated by dividing the smaller frequency count of behavior by the larger and multiplying by 100. Both observers were trained to and met a criterion of 85% agreement.

Data collection. Data were collected from teachers during the videotaping phase of the study. After every sixth day of videotaping, two teachers completed the SSRS-T independently to rate the behavior of the child being videotaped at the time. Teachers were asked to rate children's behavior displayed only during the time period of the six prior observation sessions. Thus, the first teacher rating for each child corresponded to the first six observation sessions, and the second teacher rating for each child corresponded to the second six observation sessions.

Observers coded both children's behavior from the videotapes after videotaping and training were completed. Thirty minutes of each taped play session were coded by both observers using the observation system described earlier. Observers watched and coded the videotapes in the order in which they were filmed. Coding occurred over five days, with both observers watching at the same time. Observers were able to stop the tape and view a behavior again if they wanted; however, they did not discuss behaviors as they were coding. The researcher also coded the videotapes at the same time as observers to assess reliability. After every second 30-min observation session, observers completed the

SSRS-T independently to rate the behavior of the child whose behavior was just coded from the videotape. In all, this procedure provided a total of twelve 30-min observations per observer per child, six observer ratings per observer per child, and two teacher ratings per teacher per child over the duration of the study.

Reliability checks with each observer were conducted by the researcher on 50% of the observations for each observer. The numerous reliability checks ensured there was a high level of agreement (85%) between observers, and consequently showed the observations to be reliable. Actual reliability estimates between both observers and the researcher averaged 86%.

Research Design and Data Analysis

Dependent variables in this study included the observers' and teachers' ratings on the SSRS-T and the total frequency per behavior observed on the observation system. Observations were recorded on the Social Skills Scale (i.e., Cooperation, Assertion, Self-Control subscales) and the Problem Behavior Scale. The number of assessment points provided ample assessment information to provide reliable and useful data.

The study involved a correlational design to examine the strength of the relationships within the complete observation record and between the direct observations and the SSRS ratings. Pearson correlations were calculated to test the relationships. Because the observation system used a frequency recording procedure and two different methods (i.e., observations and ratings) were used, Cohen's kappa agreement statistics could not be used to test the relationships. In addition, percentage agreement between the two methods was not useful because the rating scale and observation system did not share the same ways of counting behaviors. From Pearson correlation analyses, the researchers were able to determine how many observations were required before a representative sample of behavior was obtained and how well the data from the ratings and observations agreed. Statistical significance of correlations was tested by comparing the observed correlations to their predicted value (ρ) using Fisher's *z* transformations (p≤ .05; one-tailed).

Results

Direct observations and behavior ratings were gathered from the two observers who viewed the videotapes and ratings completed by two teachers. Interobserver agreement (i.e., agreement between the two observers' observations) averaged 86% for Child 1 and 87% for Child 2. Interrater agreement (i.e., agreement between two raters) for observers on the SSRS-T averaged 58% (*r* = .57) for Child 1 and 49% (*r*= .43) for Child 2. Finally, interrater agreement for two teachers on the SSRS-T averaged 68% (r .49) for Child 1 and 63% (*r* = .68) for Child 2. Teachers' intrarater agreement (i.e., agreement between one teacher's first and second ratings) was 68% (*r* = .61) for Teacher 1–Child 1, 78% (*r* = .76) for Teacher 2–Child 1, 90% (*r* = .93) for Teacher 1–Child 2, and 80% (*r* = .78) for Teacher 3–Child 2. Means and standard deviations of occurrences of observed behaviors are provided in Table 1.

Means and standard deviations for the total ratings of behavior frequencies on the SSRS-T and the subset of items that were both observed and rated are provided in Tables 2 and 3. Standard scores are provided to give the reader a general idea of the level of social skills and problem behaviors for each child, but were not used in any analyses: for Child 1, the teacher-rated average Social Skills standard

score on the SSRS was 88.25 (*SD* = 2.6), and his Problem Behavior standard score averaged 109.5 (*SD* = 7.8; both scores were in the average range for boys his age). For Child 2, the teacher-rated average Social Skills standard score on the SSRS was 105.5 (*SD* = 8.2), and her Problem Behavior average was 102 (*SD* = 8.2; both scores were in the average range for girls her age).

Pearson correlations that compared behavior frequencies in each observation category from the first observation to the total of behavior frequencies in each category from 12 observation sessions, the second 2 observation sessions to the total 12, the first 3 observation sessions to the total 12, and so on, were used to analyze the representativeness of observation data. A predicted correlation of .80 was used for tests of statistical significance. For Child 1, correlations with the total observation sample were greater than .80 by the first observation, but not statistically significant until the third. Therefore, for Child 1, three observation sessions were required to gather enough data to sufficiently correlate with all the observations. For Child 2, however, correlations were above .90 and statistically significant after only one observation session. The detailed statistical results for the correlations are displayed in Table 4.

Furthermore, the data obtained do not indicate that observations of problem behaviors represented the total record of observations with fewer sessions than observations of social skills. When social skills were differentiated from problem behaviors in the correlational analyses, correlations of both observations of problem behaviors and observations of social skills were above the predicted magnitude of .80 after the first day of observations.

Supplemental analyses were conducted to examine alternative ways of characterizing representativeness of observations. First, the correlations of behavior frequencies in each individual day of observation with behavior frequencies over total observation record were calculated (e.g., Observation Day 1 to Observation Days 1–12, Observation Day 2 to Days 1–12, etc.). Second, correlations were calculated to reveal how well early observations predicted later observations (e.g., Observation Day 1 to Observation Days 2–12, Observations Days 1–2 to Observation Days 3–12, Observation Days 1–3 to Observation Days 4–12, etc.).

Correlations between each individual day of observation and the total observation record were high for all four observer–child pairs. The mean result of any 1 day correlating with the total 12 days of observation for Child 1 was *r* = .93, with a range of .78 to .99 across the two observers. The mean correlation for Child 2 was *r* = .99, with a range of .96 to .99 across the two observers. Correlations predicting future observations also were strong.

Correlations for Child 1 ranged from .79 (e.g., Observation Days 1–2 correlated with Observation Days 3–12 for Observer 1) to .99 (e.g., Observation Days 1– 6 correlated with Observation Days 7–12 for Observer 1). Correlations for Child 2 were higher, all above .98 for both observers, thus indicating that early observations strongly represented later observations.

Results regarding the relationship between the observations and observers' ratings were mixed. Statistical significance was again tested against a predicted correlation of .80. The mean correlation for Child 1 after four observation sessions and two ratings was r = .44. Furthermore, only one correlation for Child 1 was above .80 (after two observations by Observer 2). For Child 2, the mean correlation after four observation sessions and two ratings was *r* = .64. Ratings by Observer 2 for Child 2 correlated with her observations highly (mean r across all ratings = .80), but this was not evidenced for Observer 1 (mean *r* = .64 across all ratings).

For all four observer–child pairs, correlations between observations and ratings did not show an increase as more observations took place; instead they fluctuated. For example, for Observer 2–Child 1 after 2 observations, $r = .84$, but after 10 observations, $r = .34$. No correlations of sufficient ($r > .80$) magnitude were obtained between the abbreviated SSRS subscale (only those items specifically observed) and observers' observations. Correlations ranged from a low of $r = .15$ for Child 1 after nine observation sessions to a high of $r = .74$ for Child 2 after 10 observations. In addition, the patterns of correlations were not consistent; that is, they fluctuated as more observations occurred.

Results of this study also addressed the correlation between the total observation record of 12 observations and teacher ratings of Total Social Skills and Problem Behavior. The magnitude of correlations was moderate between observations and teacher ratings (mean $r = .66$ across all observer–child pairs after 12th observation and 2nd teacher rating); however, no correlations reached statistical significance. Correlations generally were consistent across the entire observation record; that is, correlations between teachers' second ratings and observers' observations after 1 observation were just as high as those after 12 observations (mean $rs = .67$ and .66, respectively).

No correlations of moderate to high magnitude were obtained between the abbreviated SSRS subscale (only those items specifically observed) and observers' observations (mean $r = .18$ across all four observer–child pairs after 12 observations and two teacher ratings). Correlations for Child 1 were lower than those for Child 2 (after 12 observation sessions and two teacher ratings, mean $r = .01$ for Child 1 and mean $r = .35$ for Child 2).

In this study, note that statistical significance was tested by comparing the obtained correlations against a predicted criterion. Typically, most researchers compare correlations to a zero correlation when testing for significance. This approach is very liberal and may overstate the significance of findings. The results of our significance testing definitely would have been different (i.e., statistically significant) had the correlations been tested in this manner. All correlations between behavior ratings and observations that reached the predicted magnitude would have been statistically significant.

Discussion

The purpose of the present study was to examine the representativeness of direct observations and the validity of behavior ratings. Employing these assessment methods with preschool children in a natural setting, this research found that under optimal observation conditions (i.e., a videotaped record), up to three 30-min observation sessions were required to reliably represent a preschooler's social behavior. Conducting more than three observation sessions did not significantly improve the session intercorrelations, the indicator of the representativeness of the observations. In addition, results of this study showed that one may have confidence in behavior ratings as valid summaries of actual observed behavior. Under optimal conditions, observers' ratings of overall social behavior matched well with their own observations after only a few observations. In addition, teachers' behavior ratings of a target student matched well with observers' independent observations of that student after only a few observation sessions. Although high correlations (defined as .80) in this study between observations and ratings were not found, the cross-method correlations approach this standard. This finding suggests that neither assessment method should replace the other as a primary source of information; both methods

have unique information to add to the description of a child's behavior. Practitioners can gain useful information from both methods.

At the outset of this discussion, it is important to recognize that observations and ratings occurred with children who were not being treated for any presenting difficulties with social skills or problem behaviors. If the children had been in a treatment program and observations were being conducted to assess treatment effects, one may find that even more than 12 observations are needed to capture changes over time. The findings from this study are most relevant for professionals when they are establishing baseline or obtaining pretreatment data rather than monitoring treatment effects.

Representativeness of Observations

This study on representativeness of observations was based on Doll and Elliott's (1994) examination of the number of observation sessions required to gain a representative sample of preschoolers' social behavior. Prior to their work, only "best guess" recommendations for the amount of observations necessary to provide an adequate sample of behavior were provided in the literature (e.g., Johnston & Pennypacker, 1980; Kazdin, 1975; Shapiro, 1996). The results of the present study generally are consistent with those of Doll and Elliott, and add some noteworthy findings. Doll and Elliott found that most behaviors could be predicted after a total of 80 min of observation spread out over four sessions. By comparing cumulative observations to the total observation record of 12 days, we found that at least three observation sessions of 30 min each were adequate to reliably represent behavior with confidence that the results were not obtained due to chance alone. For Child 2, only one observation was necessary to adequately represent the total observation record for that child.

These results were replicated when representativeness was analyzed in two other ways. The additional analyses revealed that for both children, each observation was highly correlated with the total observation record, and early observations were able to predict later observations adequately. Thus, when representativeness was examined in three different ways, results indicated consistently that up to three observations were needed, but more than three observation sessions (of 30 min) generally did not add new information about the behavior being observed.

When considering these results, it is necessary to remember they were obtained under optimal observation conditions and on children not in a treatment program. Observers coded behavior from videotapes, sitting in a comfortable, quiet room without the distractions of a typical preschool classroom. They were able to stop the tape and rewind a scene if they felt they may have missed a behavior. Although representativeness of observations was established after only a few observation sessions under optimal conditions, it is very possible that observations conducted under typical conditions would take longer to achieve representativeness.

Besides creating excellent observation conditions, our videotaping method also served the purpose of providing data over real time. Because behavior was recorded every time children were present in morning free play, all behaviors exhibited by the children during unstructured play were captured on videotape. Therefore, one may be confident that the behavior observed over the 12 days was truly a large and adequate sample of overall behavior against which earlier observations could be compared.

Relationship Between Observations and Ratings

Work on the relationship between behavior observations and behavior ratings in this study was based on studies by Racine (1994) and Robertson (1993), but was designed to add to the knowledge from these two studies by identifying the point at which behavior ratings validly represent a record of observed behavior. When looking at correlations between observers' observations and their own ratings, we found that correlations between overall social skills and problem behaviors were slightly higher than the moderate correlations found by Racine. However, relative to the prediction that correlations would meet or exceed the stringent criterion of .80 by the second rating, the correlations approach this criterion.

When teacher ratings were compared to observers' observations for overall social skills and problem behaviors, correlations were higher than the low moderate correlation (r =.38) found by Robertson (1993), yet still in the moderate range. What was surprising about the teachers' reports was that the correlations between the their first rating and the first few observations were just as high as the correlations between their second rating and the last few observations. This finding indicates that the behavior rating scale seemed to be valid as both a retrospective and a prospective measure of behavior; that is, it represented observed behavior that occurred before a rating just as well as it represented or predicted observed behavior that followed the rating.

For both teachers and observers, item-to-item correlations between the specific behaviors both observed and rated were much lower than expected. There are a few possible explanations for this finding. First, the finding of lower correlations might lie in the nature of the specific behaviors being observed. Take for example, "Follows teacher directions." This behavior occurred very infrequently for both children because opportunities to follow directions generally were not provided by teachers during the unstructured play periods. Despite not seeing the behavior occur frequently, observers consistently rated the behavior as occurring very often. It appears their ratings were based on the rate at which the behavior occurred, relative to the rate at which the opportunities for it to occur were provided by teachers, rather than purely on the frequency of the behavior they saw. A second explanation for the lower item-to-item correlations is that shorter tests (i.e., rating scales, questionnaires) generally are less reliable than longer tests. The shortened version of the SSRS that was compared to the observation system is a less reliable indicator of behavior than the full version, which may have contributed to the lower correlations. Finally, an important difference between the behavior rating scale and the observation format is the restricted range of the ratings. Ratings were based on a 3-point Likert-type scale whereas frequency counts of behavior were not restricted to a scale. This difference in the possible range in scores, in particular the restricted range of ratings, could be yet another explanation of the moderate correlations consistently found between ratings and observations.

Issues of Generalization

Three questions about the generalization of the findings of this study deserve attention. First, would these results occur if observations were conducted in the classroom instead of on videotape? Second, would other observers produce similar results? Finally, would these results be consistent for other children?

We believe the results are generalizable to observations directly in the classroom. Videotaping focused directly on one child for a continuous period of time. The camera recorded exactly what a live observer could have seen. Taping stopped when the child was out of sight, just as behavior observations would stop when a child goes out of view. The main difference between live observations and the observations

from videotapes in this study was the ability for observers watching on videotape to stop the tape, rewind, and view behaviors again if they desired (which happened about once per observation session). These observation conditions may have created results better than those that would be found in naturalistic observations. In fact, correlations of representativeness of observations obtained by Doll and Elliott (1994), who used live, in-class observations, were slightly lower than those found in the present study. Nevertheless, the overall pattern of results was similar, indicating the observational results from videotapes are likely to generalize to in-vivo observations, assuming the observer is well trained and uses a well-designed observation system.

The question of whether other observers would produce similar results is difficult to answer, given that there were only two observers in this study; however, the observers proved their reliability consistently and produced similar results for each child they observed. Therefore, one may conclude that with sufficient training in the observation system, other observers are likely to exhibit the same degree of reliability and produce similar results.

The question of whether the results are generalizable to other children also is difficult to answer because there were only two preschool children observed. Although one child had a diagnosed disability (Down's syndrome), the children were relatively similar in the behaviors they exhibited. It is questionable whether results would be similar with children exhibiting a high degree of problem behaviors, for example. Again, one may look to Doll and Elliott (1994) for a possible answer to this question. They found that readily noticed behaviors, such as physical aggression, were easier to predict than less noticeable social behaviors, indicating that the results may not be similar for children who exhibit different kinds of behaviors than the children in the present study. Would the results generalize to older children? Preschool-age children typically are less consistent in their behaviors than older children (Bracken, 1991). Therefore, it may be possible that fewer than three observations would be necessary to gain a reliable sample of behavior for school-age children; however, this is only speculation, and worthy of future research.

An important component of the research design of this study was its replication features. Although only two children were observed by two observers, observations and ratings were completed by multiple people over multiple days, providing replication of results across people, time, and assessment measures as well as within observers, teachers, and children. Results generally were consistent between observers, teachers, and children for all data points. In addition, results within each child were consistent across time. This replication of results further increases the likelihood of generalization to other children and observers.

Implications of the Findings

The findings of this study have implications for researchers and practitioners. For researchers, especially those developing assessment tools, these findings shed light on the validation process for a rating scale or observation system. An important component in test development is providing data on the validity of a measure, yet specific research comparing observations to related ratings for a given measure often is not found. This study offers a paradigm for individuals beginning to validate observation systems and rating scales that are related. Its findings indicate that even when measures are related, the relationship between the two may only be moderate at best.

The implications of the findings for practitioners are twofold. First, to increase confidence in the representativeness of one's observations, the findings of this study indicate at least three 30-min observations should be conducted. More sessions may be needed depending on the purpose of the observations and the child being observed. Second, even under excellent observation conditions, only moderate correlations were found between observations and ratings by both teachers and observers. This finding indicates that neither method should replace the other as a source of information; they are providing some similar and some different information. Thus, it seems practitioners would do best to employ a multisource, multimethod assessment as consistently suggested in the best practices literature (e.g., Harrison & Robinson, 1995; Knoff, 1995; Landau & Burchman, 1995; McConaughy & Ritter, 1995).

The use of videotaped observation in this study also has an implication for training. School psychology students, practitioners, and parents can use the recorded behavior to learn how to complete structured observations and behavior ratings. The use of the videotapes as a tool for learning and practice could enhance the trainees' reliability as sources of assessment.

Suggestions for Future Research

While videotaping of children was carried out to create a naturalistic observation environment, observers still had the opportunity to stop the videotapes to review behaviors, creating optimal observation conditions. It may be useful in future research to have observers view behavior from the videotapes *without* stopping or to perform observations directly in classrooms. This additional research would increase the confidence that the results obtained in the present study would generalize to typical practice in which observations are not usually conducted under optimal conditions.

In addition, it would be useful for future research to have increased sample sizes with students who vary in age in addition to their behaviors. The two children in the present study may represent children with and without disabilities in their classroom, but children of other ages, abilities, and behaviors may or may not provide the same results. Research with different observers also would be necessary to ensure the generalization of the results.

Conclusions

Recall that this study was motivated by two practical questions: (a) How many observation sessions are needed before a reliable and valid assessment of children's social behavior is obtained? (b) How much confidence can one have that a well-developed rating scale adequately reflects actual observed behaviors? The evidence collected and synthesized in this study provides some answers to these questions.

First, it was demonstrated that multiple observation sessions of preschoolers' social behavior are needed to gain a sample of behavior that represents the children's overall behavior. Furthermore, behavior ratings do an adequate job of representing observed behavior when teachers and observers rate behaviors. Practitioners who regularly use observations and behavior ratings as part of a multisource, multimethod assessment practice have a responsibility to use methods that are reliable and valid. The results of this study demonstrate what it takes for practitioners to use these methods with confidence.

Second, the present study provided a foundation for future research in the area of representativeness and validity of observations and ratings. Further research with an increased sample size, older children,

children with more variability in behaviors, and in-class observations would be useful to have confidence that the results would generalize to other populations. Research with an incontrovertible index of behavior and with rating scales that use a wider range of ratings also would be useful.

The results of this study provide useful information that contributes both practical and theoretical knowledge of assessment of children's social behavior. Practitioners and researchers interested in a representative and reliable picture of young students' behavior are encouraged to use rating scales in conjunction with observation methods that sample behavior at least three points in time.

Table 1
Means and Standard Deviations of Daily Frequencies[a] of Observed Behaviors Over Complete Observation Record per Observer/Child Pair

Behavior	Observer–Child Pair			
	Observer 1–Child 1	Observer 2–Child 1	Observer 1–Child 2	Observer 2–Child 2
Cooperation	**17.33**	**17.83**	**16.00**	**15.50**
	(3.55)	**(3.90)**	**(1.95)**	**(1.31)**
Joins activity	1.75	1.00	2.25	1.58
	(2.53)	(1.35)	(1.76)	(1.31)
Participating	12.58	12.58	13.75	13.58
	(2.54)	(2.75)	(1.14)	(1.38)
Follow directions	3.00	4.25	0.00	0.33
	(2.13)	(3.55)	(0.00)	(0.65)
Assertion	**0.67**	**0.25**	**0.25**	**0.33**
	(0.98)	**(0.45)**	**(0.45)**	**(0.49)**
Invites others	0.67	0.25	0.17	0.17
	(0.98)	(0.45)	(0.39)	(0.39)
Initiates conversations	0.00	0.00	0.08	0.17
	(0.00)	(0.00)	(0.29)	(0.39)
Self-Control	**3.42**	**3.50**	**12.50**	**12.67**
	(2.31)	**(2.88)**	**(1.78)**	**(1.56)**
Controls temper	0.00	0.00	0.25	0.33
	(0.00)	(0.00)	(0.00)	(0.65)
Waits turn	1.50	1.08	0.00	0.00
	(2.58)	(1.56)	(0.00)	(0.00)
Cooperates without prompting	1.92	2.42	12.25	12.33
	(2.15)	(2.75)	(1.86)	(1.83)
Social Skills Total	**21.42**	**21.58**	**28.75**	**28.50**
	(4.83)	**(4.56)**	**(3.33)**	**(2.43)**
Aggressive	0.00	0.08	0.17	0.08
	(0.00)	(0.29)	(0.39)	(0.29)
Argues	0.25	0.08	0.33	0.33
	(0.62)	(0.29)	(0.65)	(0.89)
Acts sad or depressed	0.08	0.08	0.00	0.00
	(0.29)	(0.29)	(0.00)	(0.00)
Problem Behaviors Total	**0.33**	**0.25**	**0.50**	**0.42**
	(0.65)	**(0.45)**	**(0.80)**	**(0.90)**

Note. Standard deviations are in parentheses below means. "Social Skills Total" is made up of all the Cooperation, Assertion, and Self-Control items in the observation system.

[a]Frequencies of observed behaviors refers to the number of occurrences of behaviors, not the rate or proportion of occurrence.

Table 2
Means and Standard Deviations of Observers' Behavior Ratings Over Complete Rating Period[a] *per Observer–Child Pair*

Behavior	Observer–Child Pair			
	Observer 1–Child 1	Observer 2–Child 1	Observer 1–Child 2	Observer 2–Child 2
Cooperation	**11.67**	**11.83**	**15.83**	**15.67**
	(2.66)	**(1.17)**	**(2.40)**	**(1.51)**
Joins activity	0.67	1.00	1.83	1.17
	(0.82)	(0.00)	(0.41)	(0.41)
Participating	2.00	1.33	2.00	1.33
	(0.00)	(0.52)	(0.00)	(0.52)
Follow directions	1.83	2.00	2.00	2.00
	(0.41)	(0.00)	(0.00)	(0.00)
Assertion	**7.83**	**6.00**	**14.00**	**10.50**
	(1.94)	**(2.45)**	**(3.85)**	**(3.27)**
Invites others	1.67	1.17	0.83	1.00
	(0.52)	(0.41)	(0.75)	(0.00)
Initiates conversations	1.00	0.67	1.00	1.00
	(0.89)	(0.52)	(0.63)	(0.00)
Self-Control	**11.67**	**15.50**	**13.17**	**16.67**
	(2.73)	**(2.35)**	**(1.72)**	**(2.58)**
Controls temper	0.83	1.83	1.17	1.33
	(0.41)	(0.41)	(0.41)	(0.52)
Waits turn	2.00	1.83	1.50	1.83
	(0.00)	(0.41)	(0.55)	(0.41)
Cooperates without prompting	1.00	1.33	2.00	1.83
	(0.89)	(0.52)	(0.00)	(0.41)
Social Skills Total	**31.17**	**33.17**	**43.00**	**42.83**
	(6.62)	**(5.19)**	**(7.62)**	**(6.82)**
Aggressive	0.00	0.00	1.00	0.17
	(0.00)	(0.00)	(0.00)	(0.41)
Argues	0.17	0.33	0.83	0.17
	(0.41)	(0.52)	(0.41)	(0.41)
Acts sad or depressed	0.50	0.17	0.50	0.17
	(0.84)	(0.41)	(0.55)	(0.41)
Problem Behaviors Total	**4.00**	**2.83**	**6.00**	**1.00**
	(1.10)	**(0.75)**	**(2.10)**	**(1.10)**

Note. Standard deviations are in parentheses below the means. Cooperation is made up of all Cooperation items on the SSRS-T; Assertion is made up of all Assertion items on the SSRS-T; Self-Control is made up of all Self-Control items on the SSRS-T; Social Skills Total is made up of all Cooperation, Assertion, and Self-Control items on the SSRS-T; Problem Behaviors is made up of all Problem Behavior items on the SSRS-T.

[a]Complete rating period is six ratings per observer per child.

Table 3
Means and Standard Deviations of Teachers' Behavior Ratings Over Complete Rating Period[a] *per Teacher–Child Pair*

Behavior	Teacher 1–Child 1	Teacher 2–Child 1	Teacher 1–Child 2	Teacher 3–Child 2
Cooperation	**10.50**	**10.00**	**18.00**	**16.00**
	(0.71)	**(1.41)**	**(1.41)**	**(0.00)**
Joins activity	0.71	0.00	2.00	2.00
	(0.00)	(0.00)	(0.00)	(0.00)
Participating	1.00	1.00	2.00	2.00
	(0.00)	(0.00)	(0.00)	(0.00)
Follow directions	1.00	1.00	2.00	2.00
	(0.00)	(0.00)	(0.00)	(0.00)
Assertion	**3.50**	**5.50**	**15.00**	**13.50**
	(0.71)	**(0.71)**	**(1.41)**	**(0.71)**
Invites others	1.00	0.00	1.50	1.00
	(0.00)	(0.00)	(0.71)	(0.00)
Initiates conversations	0.50	1.00	2.00	2.00
	(0.71)	(0.00)	(0.00)	(0.00)
Self-Control	**11.00**	**12.00**	**18.00**	**14.00**
	(0.00)	**(1.41)**	**(0.00)**	**(0.00)**
Controls temper	2.00	1.00	2.00	1.50
	(0.00)	(0.00)	(0.00)	(0.71)
Waits turn	2.00	2.00	2.00	2.00
	(0.00)	(0.00)	(0.00)	(0.00)
Cooperates without prompting	1.00	1.00	2.00	2.00
	(0.00)	(0.00)	(0.00)	(0.00)
Social Skills Total	**25.00**	**27.50**	**51.00**	**43.50**
	(0.00)	**(0.71)**	**(2.83)**	**(0.71)**
Aggressive	1.00	1.00	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)
Argues	1.00	1.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Acts sad or depressed	0.50	1.00	0.00	1.00
	(0.71)	(0.00)	(0.00)	(0.00)
Problem Behaviors Total	**6.00**	**10.00**	**2.00**	**5.50**
	(1.41)	**(1.41)**	**(0.00)**	**(0.71)**

Note. Standard deviations are in parentheses below the means. Cooperation is made up of all Cooperation items on the SSRS-T; Assertion is made up of all Assertion items on the SSRS-T; Self-Control is made up of all Self-Control items on the SSRS-T; Social Skills Total is made up of all Cooperation, Assertion, and Self-Control items on the SSRS-T; Problem Behaviors is made up of all Problem Behavior items on the SSRS-T.

[a]Complete rating period is two ratings per teacher per child.

Table 4
Pearson Correlations Between Total Observation Record and Cumulative Days of Observation per Observer–Child Pair

	Total Observation Record			
Day(s) of observation	Observer 1–Child 1	Observer 2–Child 1	Observer 1–Child 2	Observer 2–Child 2
1	.89	.87	.98*	.99*
1–2	.84	.90	1.00*	.99*
1–3	.96*	.97*	1.00*	.99*
1–4	.98*	.98*	1.00*	.99*
1–5	.99*	.99*	1.00*	1.00*
1–6	1.00*	1.00*	1.00*	1.00*
1–7	1.00*	1.00*	1.00*	1.00*
1–8	1.00*	1.00*	1.00*	1.00*
1–9	1.00*	1.00*	1.00*	1.00*
1–10	1.00*	1.00*	1.00*	1.00*
1–11	1.00*	1.00*	1.00*	1.00*
Total observation Record (1–12)	1.00*	1.00*	1.00*	1.00*

*$p < .05$ (one-tailed) for ρ (rho) > .70.

References

Achenbach, T., & Edelbrock, C. (1986). Manual for the teacher's report form and teacher version of the child behavior profile. Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T.M., McConaughy, S.H., & Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin, 101, 213–232.

Barton, E.J., &Ascione, F.R. (1984). Direct observation. In T. Ollendick &M. Hersen (Eds.), Child behavioral assessment: Principles and procedures (pp. 166–194). New York: Pergamon Press.

Bracken, B.A. (1991). The clinical observation of preschool assessment behavior. In B. Bracken (Ed.), The psychoeducational assessment of preschool children (pp. 40–52). Boston: Allyn & Bacon.

Demaray, M.K., Ruffalo, S.L., Carlson, J., Busse, R.T., Olson, A.E., McManus, S.M., & Leventhal, A. (1995). Social skills assessment: A comparative evaluation of six published rating scales. School Psychology Review, 24, 648– 671.

Doll, B., & Elliott, S.N. (1994). Representativeness of observed preschool social behaviors: How many data are enough? Journal of Early Intervention, 18, 227–238.

Elliott, S.N., Gresham, F.M., Freeman, T., & McCloskey, G. (1988). Teacher and observer ratings of children's social skills: Validation of the social skills rating scales. Journal of Psychoeducational Assessment, 6, 152–161.

Foster, S.L., & Cone, J.D. (1995). Validity issues in clinical assessment. Psychological Assessment, 7, 248–260.

Gresham, F.M., & Elliott, S.N. (1990). Social skills rating system. Circle Pines, MN: American Guidance Service.

Harrison, P.L., & Robinson, B. (1995). Assessment of adaptive behavior. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology III (pp. 753–762). Washington, DC: National Association of School Psychologists.

Johnston, J.M., & Pennypacker, H.S. (1980). Strategies and tactics of human behavioral research. Hillsdale, NJ: Erlbaum.

Kazdin, A.E. (1975). Behavior modification in applied settings. Homewood, IL: Dorsey Press.

Knoff, H.M. (1995). Personality assessment. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology III (pp. 849–864). Washington, DC: National Association of School Psychologists.

Landau, S., & Burchman, B.G. (1995). Assessment of children with attention disorders. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology III (pp. 817–830).Washington, DC: National Association of School Psychologists.

McConaughy, S.H., & Ritter, D.R. (1995). Multidimensional assessment of emotional or behavioral disorders. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology III (pp. 865–877). Washington, DC: National Association of School Psychologists.

Merrell, K.W. (1993a). School social behavior scales. Bradon, Austin, TX: Pro-Ed.

Merrell, K.W. (1993b). Using behavior rating scales to assess social skills and antisocial behavior in school settings: Development of the school social behavior scales. School Psychology Review, 22, 115–133.

Racine, C.N. (1994). The relationship between observations and ratings of children's social behavior: An extension of the accuracy–reliability paradigm. Unpublished master's thesis, University of Wisconsin, Madison.

Robertson, S.L. (1993). The relationship between observations and ratings of children's social behavior. Unpublished master's thesis, University of Wisconsin, Madison.

Shapiro, E.S. (1996). Academic skills problems: Direct assessment and intervention. New York: Guilford Press.