Nonexpert Ratings of Family and Parent–Child Interaction

Jason K. Baker
University of Wisconsin–Madison

Daniel S. Messinger, Naomi V. Ekas, Kristin M. Lindahl, and Ryan Brewster
University of Miami

Observational methods benefit the study of family process, but many expert rating systems are costly and time-consuming. This study examined the utility of using small groups of eight to ten nonexperts to rate family conflict and maternal sensitivity. Videotaped triadic interactions of 39 families were drawn from Lindahl (1998), and 22 mother–toddler free-play interactions were drawn from Baker, Messinger, Lyons, and Grantz (2010). Sixty undergraduates rated interactions from these samples in real time using computer-assisted technology. Nonexpert ratings of family conflict were reliable, demonstrated high concordance with expert ratings, and replicated a key finding from Lindahl (1998). Nonexpert ratings of maternal sensitivity replicated a relevant finding from Baker, Messinger et al. (2010). Concordance was lower for maternal sensitivity, however, because of the tendency of nonexperts to overattend to sensitive structuring compared with emotional supportiveness. A second study indicated that as few as six nonexperts could effectively rate maternal sensitive structuring, but that nonexperts were unable to accurately rate emotional supportiveness. Implications for research methods and for our understanding of these important family constructs are discussed.

Keywords: methods, sensitivity, family conflict, nonexperts, observation

Family-focused research typically utilizes either questionnaires or direct observation to measure interactive behavior. There are important limitations to full reliance upon questionnaire data, including reporter bias and difficulty measuring more complicated and subtle behavior. On the other hand, many observational systems designed to capture family interaction are complex, costly, and time-consuming, placing limits on the use of this valuable research approach. The present study considered the utility of nonexpert ratings to the study of family process.

Investigations of family functioning often focus on either the family system or on a particular dyad within the family.

Lindahl (1998) found that families of children with oppositional defiant disorder (ODD) exhibited more conflict and rejection-coercion than did control families. These data were obtained using the System for Coding Interactions and Family Functioning (SCIFF; Lindahl & Malik, 1996), which required 15 hours of initial training, continuous monitoring, and weekly feedback meetings (Lindahl, 1998). Raters watched each family a minimum of three times, and rating a 10-min interaction required 35 to 45 min (Lindahl & Malik, 2001). Similar or more extensive effort has been reported for other systems focused on family-level dynamics (see Kerg & Lindahl, 2001).

Rating systems that assess dyadic parent–child interaction are similarly complex and often require considerable training and rating time. Some require training by the developers of the system at special workshops, followed by a certification/reliability process (e.g., Biringen, Robinson, & Emde, 1998; NICHD Early Child Care Research Network, 1996). Other systems require rigorous training lasting up to 6 months before raters begin to process data (e.g., Aber, Belsky, Slade, & Crnic, 1999). The maternal sensitivity scales of the National Institute of Child Health and Human Development Early Child Care Research Network (NICHD Early Child Care Research Network, 1999) assess mothers’ emotional tone, responsiveness, respect for child autonomy, and ability to structure the child. A recent study by Baker, Messinger, Lyons, & Grantz (2010) found that these sensitivity ratings related to fewer behavior problems in a subset.
of their sample of toddlers. Training for this study was similar to that reported by Aber et al. (1999) and included several months (at 6–8 hrs/wk) of initial training, followed by continuous monitoring and weekly feedback meetings. Pairs of bachelor’s-level research assistants rated approximately three 5-min interactions per hour.

Most individuals involved in rating family interaction possess considerable relevant personal experience. The substantial time and effort devoted to training is often focused on reducing individual interpretive biases and increasing adherence to objective guidelines. However, important information may lie in the personal experiences of observers. Indeed, nonexpert ratings can contribute to the field of observational measurement (e.g., Gottman & Levenson, 1985), with the predictive ability of these ratings leading some to assert that nonexperts possess “intuitive expertise” in certain domains (Waldinger, Schulz, Hauser, Allen, & Crowell, 2004). One way to harness the benefit of shared personal experience while reducing individual bias is to aggregate scores from multiple raters (Larrick & Soll, 2004). One way to harness the benefit of shared personal experience while reducing individual bias is to aggregate scores from multiple raters (Larrick & Soll, 2006). Waldinger et al. (2004) utilized this so-called wisdom of crowds phenomenon (Surowiecki, 2004) and found that ratings of marital interaction averaged across a small group of nonexperts were highly correlated with expert ratings. Baker, Baltigan, Brewster, Jaccard, and Messinger (2010) found concordance between expert codes of infant facial actions and mean ratings of infant emotion produced by 5 to 10 nonexperts. Furthermore, nonexpert data resulted in findings similar to those derived from the expert codes. Importantly, concordance for parent emotion ratings was lower, suggesting that more complex behavior might prove challenging for nonexperts.

The current study examined the ability of nonexperts to rate family conflict and maternal sensitivity—the former requiring attention to the interdependent behavior of multiple individuals, and the latter emphasizing consideration of multiple behavioral modalities. Study 1 examined concordance between novel nonexpert ratings and subsets of existing expert data from Lindahl (1998) and Baker, Messinger et al. (2010). Nonexpert data were then used to replicate relevant findings from these two investigations. Finally, Study 2 involved data collection and analysis to clarify how nonexperts perceived sensitivity in Study 1.

Study 1

Methods

Participants. Videotaped segments rated by nonexperts included 39 of the original Lindahl (1998) interactions and 22 interactions from Baker, Messinger, et al. (2010). Lindahl (1998) included 110 multi-ethnic, two-parent families of boys (current sample, $M_{age} = 8.69$, $SD = 1.22$), and Baker, Messinger, et al. (2010) included a culturally diverse group of 35 mothers and their 18-month-olds with and without early autism risk. Families were randomly selected, except that children with only attention-deficit/hyperactivity disorder in Lindahl (1998) were undersampled because conflict did not discriminate this group (current $n = 6$) from the controls. Although our concordance analyses utilized all selected families, replication analyses were limited to the Lindahl (1998) families from the control ($n = 18$) and ODD groups ($n = 15$), and the Baker, Messinger, et al. (2010) children without autism diagnoses (those for whom the negative relation between sensitivity and behavior problems was observed; $n = 12$).

Nonexperts were 60 students recruited from a university community in the southeastern United States. Although some students were psychology majors, most were attending their first psychology course, and all reported that they had not completed coursework directly relevant to the particular construct. Forty nonexperts rated conflict ($M_{age} = 19.41$ years, $SD = 1.25$; 62% women; 47% European American) and 20 rated sensitivity ($M_{age} = 19.06$ years, $SD = 1.56$; 53% women; 77% European American). Nonexperts individually rated 9 to 12 families, and each family was rated by 10 nonexperts.

Procedures. Procedures for the current study and the two original studies were performed in accordance with the relevant institutional review boards. Details regarding families, protocols, and expert rating can be found in the original investigations. Briefly, families in Lindahl (1998) were asked to discuss a recent argument that included both parents and the son. Mothers from Baker, Messinger, et al. (2010) were given toys and asked to play with their toddlers as they would at home.

After consent was obtained, nonexperts were familiarized with a joystick and digitized video interface. Family conflict raters were asked to “rate how much conflict occurs in the family, including how coercive and rejecting the parents are. Conflict also refers to any tension and negativity in the interaction.” Sensitivity raters were told to rate “the degree to which a parent appears tuned into their child’s needs and engages in warm, supportive behaviors that follow the child’s lead when possible and provide structure for the child when appropriate.” The Continuous Measurement System (CMS; http://measurement.psy.miami.edu/cms.phtml; Messinger, Mahoor, Chow, & Cohn, 2009) presented video files to the nonexperts in a random sequence and raters used a joystick to indicate how much of the construct in question was being exhibited. Nonexperts rated in real-time such that the rating time equaled the length of each interaction (12 min for each family discussion, 5 min for each free play). The CMS recorded rating data for each frame of the video (30 per second).

Measures. Expert ratings of family conflict and maternal sensitivity were based on the SCIFF (Lindahl & Malik, 1996) and the NICHD Early Child Care Research Network scales (1999), respectively. The expert conflict rating was obtained by averaging the SCIFF scores of rejection-coercion and conflict (the ratings that characterized the ODD group). The expert rating of sensitivity included the average of emotional supportiveness (responsiveness, respect for autonomy, positive regard) and sensitive structuring. Each nonexpert’s rating data for each family were averaged across the video frames. Family scores were then averaged across raters. Replication of the original findings...

Results

Reliability and concordance with expert scores. Average-measures intraclass correlations (ICCs) were conducted for the nonexpert ratings (see Baker, Haltigan et al., 2010). Average reliability for the conflict ratings was acceptable at .76, and concordance between expert and nonexpert ratings was high, $r = .81$, $p < .001$. Follow-up analyses revealed that as few as eight raters could provide reliable ratings that corresponded well with expert ratings, ICC = .71; $r = .80$, $p < .001$. As would be expected from truncating the variance, concordance was somewhat lower when only ratings of the control families were considered, but remained moderate, $r = .69$, $p < .01$.

Reliability for maternal sensitivity was high, ICC = .85, but concordance with expert ratings was moderate, $r = .65$, $p < .01$. Eight raters were necessary for acceptable reliability, ICC = .76, and concordance was unchanged, $r = .65$, $p < .01$. A post-hoc hypothesis posited that nonexperts overattended to maternal teaching compared with more affective aspects of sensitivity. Indeed, subsequent analyses indicated that sensitivity ratings from eight nonexperts corresponded well with the original expert ratings of sensitive structuring, $r = .80$, $p < .001$, and were unrelated to expert emotional supportiveness ratings, $r = .39$, $p < .10$.

Replication of findings from Lindahl (1998) and Baker, Messinger, et al. (2010). Ratings derived from eight nonexperts replicated the relevant finding from Lindahl (1998). Family conflict was higher in families of children with ODD ($M = 80.97, SD = 84.56$) than those without ODD ($M = 30.73, SD = 80.19$), controlling for family income, $F = 10.98, p < .01$, $\eta^2_p = .27$. Sensitivity ratings from eight nonexperts replicated the finding from Baker, Messinger, et al. (2010), in that sensitivity was related to behavior problems in children without eventual autism, $r_s = -.64$, $p < .05$.

Study 2

Given that nonexperts appeared to rate structuring when asked to rate sensitivity, Study 2 tested whether nonexperts could effectively rate the sub-components of sensitivity. Two sets of 10 additional nonexperts ($M_{age} = 19.61$ years; $SD = 1.20$; 69% women; 25% European American) each rated the same set of 12 Baker, Messinger et al. (2010) families. One set rated sensitive structuring (“the degree to which the parent is involved in providing appropriate structure and teaching for the child”), and one set rated emotional supportiveness (“the degree to which the parent is warm, positive, responsive and supportive to her child, while also respecting the child’s independence”). As few as six nonexperts provided reliable rating of structuring, ICC = .75, and concordance with expert ratings was moderately high, $r = .71$, $p < .001$. Emotional supportiveness ratings, however, were neither reliable nor associated with expert ratings, ICC = .47, $r = .36$, ns. In fact, these nonexpert emotional supportiveness scores exhibited moderately high associations with the structuring ratings of experts, $r = .78$, $p < .001$, and nonexperts, $r = .70$, $p < .01$.

General Discussion

Findings suggest that small groups of nonexperts can effectively rate certain aspects of parenting and family process. Aggregate nonexpert ratings of family conflict were reliable, concordant with expert scores, and replicated a relevant finding from Lindahl (1998). Nonexperts were able to adequately recognize maternal sensitive structuring, and nonexpert sensitivity ratings replicated findings from Baker, Messinger, et al. (2010). However, nonexperts were not effective in rating mothers’ emotional supportiveness.

The ability of nonexperts to recognize family conflict likely stems from shared personal and cultural experiences. Although individual differences in perceptions of conflict were present, aggregating ratings over a small group reduced individual bias and harnessed common experience. Nonexperts appear similarly attuned to the way in which mothers sensitively teach their children; in fact, they may overattend to this behavior in their notions of sensitivity. Emotional supportiveness is seemingly more complex and multi-faceted than is structuring, thus it may be more difficult to rate even when raters are explicitly told to attend to certain behaviors. Nonexpert sensitivity ratings nonetheless replicated the relevant finding from Baker, Messinger et al. (2010), likely because structuring drove this original finding.

Nonexperts in the present study were “trained” with one to two sentences and rated each segment in real time. However, this method only produced a single score, whereas most relevant expert systems generate several ratings. Thus, nonexpert ratings with the CMS system are best suited for assessment of a single construct that can be rated on a continuum. The family conflict rating was also relatively broad, encompassing several elements of problematic functioning. Nonexperts may not prove as adept at discriminating between more specific negative behaviors such as conflict and coercion—an important distinction given the differential importance of these constructs to outcomes such as antisocial behavior (Granic & Patterson, 2006). Although our families were normative on the sensitivity dimension (see Baker, Messinger, et al., 2010), Lindahl (1998) included families with more extreme conflict, and concordance in the present study was somewhat lower when only control families were considered. It is likely that nonexperts adjusted their range of potential scores based upon the full range of observed behavior, thus the variation among more moderate scores was truncated for the post hoc examination.

Our method of grouping nonexperts and the relatively small number of raters precluded consideration of nonexpert ethnicity as a moderating factor. Future studies should include a larger, culturally-diverse group of nonexperts and
should explore associations between additional characteristics (e.g., being a parent) and individuals’ ability to rate certain constructs. Messinger, Cassel, Acosta, Ambadar, & Cohn (2008) found that parent and undergraduate ratings of infant affect were essentially identical, but parents might provide more nuanced rating of sensitivity than undergraduates. It is possible that nonexpert ratings could be improved if slightly more time (e.g., 10 min) were devoted to “training.” The present study utilized only the mean rating of each interaction; however, a key strength of continuous measurement is its ability to provide data over time (Chow, Haltigan, & Messinger, 2010). Indeed, it is possible that the real-time method used by nonexperts may have provided an advantage over the expert molar approach, and future studies examining continuous expert coding would be interesting. Finally, some constructs appear more recognizable than others, and exploring which aspects of family functioning may be more or less amenable to nonexpert rating is a fruitful avenue for future research, one that will likely provide valuable information about the constructs we investigate.

References


Received May 5, 2010
Revision received July 28, 2010
Accepted August 16, 2010