

Non-Expert Ratings of Family and Parent-Child Interaction

Jason K. Baker,
Waisman Center, University of Wisconsin-Madison

Daniel S. Messinger,
University of Miami

Naomi V. Ekas,
University of Miami

Kristin M. Lindahl, and
University of Miami

Ryan Brewster
University of Miami

Abstract

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 non-experts 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. Non-expert ratings of family conflict were reliable, demonstrated high concordance with expert ratings, and replicated a key finding from Lindahl (1998). Non-expert ratings of maternal sensitivity replicated a relevant finding from Baker, et al. (2010). Concordance was lower for maternal sensitivity, however, due to the tendency of non-experts to over-attend to sensitive structuring as compared to emotional supportiveness. A second study indicated that as few as six non-experts could effectively rate maternal sensitive structuring, but that non-experts 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; non-experts; 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 non-expert 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-minute interaction required 35 to 45 minutes (Lindahl & Malik, 2001). Similar or more extensive effort has been reported for other systems focused on family-level dynamics (see Kerig & 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 ECCRN, 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 ECCRN, 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-minute 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, non-expert 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 non-experts 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, 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 non-experts were highly correlated with expert ratings. Baker, Haltigan, 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 non-experts. Furthermore, non-expert 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 non-experts.

The current study examined the ability of non-experts 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 non-expert ratings and subsets of existing expert data from Lindahl (1998) and Baker, Messinger et al. (2010). Non-expert data were then used to replicate relevant findings from these two investigations. Finally, Study 2 involved data collection and analysis in order to clarify how non-experts perceived sensitivity in Study 1.

Study 1

Method

Participants—Videotaped segments rated by non-experts 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 $M_{\text{age}} = 8.69$, $SD = 1.22$), and Baker, Messinger et al. (2010) included a culturally-diverse group of 33 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 under-sampled 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; $n = 12$).

Non-experts 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 non-experts rated conflict ($M_{\text{age}} = 19.41$ years, $SD = 1.25$; 62% female; 47% European American) and 20 rated sensitivity ($M_{\text{age}} = 19.06$ years, $SD = 1.56$; 53% female; 77% European American). Non-experts individually rated 9 to 12 families, and each family was rated by 10 non-experts.

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, non-experts 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 non-experts in a random sequence and raters used a joystick to indicate how much of the construct in question was being exhibited. Non-experts rated in real-time such that the rating time equaled the length of each interaction (12 minutes for each family discussion, 5 minutes 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 ECCRN 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 non-expert’s rating data for each family were averaged across the video frames. Family scores were then averaged across raters. Replication of the original findings involved two additional measures. Lindahl assigned

boys to the ODD group based on the ODD Rating Scale (Lindahl, 1998). Baker, Messinger et al. (2010) examined relations between sensitivity and parent report on the Child Behavior Checklist (Achenbach, 2000), a measure of child behavior problems.

Results

Reliability and concordance with expert scores—Average-measures intra-class correlations (*ICC*) were conducted for the non-expert ratings (see Baker, Haltigan et al., 2010). Average reliability for the conflict ratings was acceptable at .76, and concordance between expert and non-expert 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 non-experts over-attended to maternal teaching as compared to more affective aspects of sensitivity. Indeed, subsequent analyses indicated that sensitivity ratings from eight non-experts 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 non-experts 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, $F = 10.98, p < .01, \eta_p^2 = .27$. Sensitivity ratings from eight non-experts 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 non-experts appeared to rate structuring when asked to rate sensitivity, Study 2 tested whether non-experts could effectively rate the sub-components of sensitivity. Two sets of 10 additional non-experts ($M_{age} = 19.61$ years; $SD = 1.20$; 69% female; 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 non-experts 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 non-expert emotional supportiveness scores exhibited moderately high associations with the structuring ratings of experts, $r = .78, p < .001$, and non-experts, $r = .70, p < .01$.

General Discussion

Findings suggest that small groups of non-experts can effectively rate certain aspects of parenting and family process. Aggregate non-expert ratings of family conflict were reliable, concordant with expert scores, and replicated a relevant finding from Lindahl (1998). Non-experts were able to adequately recognize maternal sensitive structuring, and non-expert

sensitivity ratings replicated findings from Baker, Messinger et al. (2010). However, non-experts were not effective in rating mothers' emotional supportiveness.

The ability of non-experts 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. Non-experts appear similarly attuned to the way in which mothers sensitively teach their children; in fact, they may over-attend 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. Non-expert sensitivity ratings nonetheless replicated the relevant finding from Baker, Messinger et al. (2010), likely because structuring drove this original finding.

Non-experts 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, non-expert 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. Non-experts 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 non-experts 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 non-experts and the relatively small number of raters precluded consideration of non-expert ethnicity as a moderating factor. Future studies should include a larger, culturally-diverse group of non-experts 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 non-expert ratings could be improved if slightly more time (e.g., 10 minutes) 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 non-experts 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 non-expert rating is a fruitful avenue for future research, one that will likely provide valuable information about the constructs we investigate.

Acknowledgments

The Continuous Measurement System can be downloaded free of charge at <http://measurement.psy.miami.edu/cms.phtml>. This study was funded by NIH grants R21HD052062 and R01HD047417 (PI, Messinger), T32 HD007473 (University of Miami), and T32 HD07489 (University of Wisconsin-Madison). We thank Seniz Celimli, Thomas Oswald, Kristina Astone, and staff of the UM Sib Study for assistance with data collection. Ryan Brewster is now at Georgia State University.

References

- Aber J, Belsky J, Slade A, Crnic K. Stability and change in mothers' relationship representations with their toddlers. *Developmental Psychology* 1999;35:1038–1047. [PubMed: 10442872]
- Achenbach, TM. *Manual for the Child Behavior Checklist 1 ½ – 5*. Burlington: University of Vermont, Department of Psychiatry; 2000.
- Baker JK, Haltigan J, Brewster R, Jaccard J, Messinger D. Non-expert ratings of infant and parent emotion: Concordance with expert coding and relevance to early autism risk. *International Journal of Behavioral Development* 2010;34:88–95.
- Baker JK, Messinger DS, Lyons KK, Grantz CJ. A pilot study of maternal sensitivity in the context of emergent autism. *Journal of Autism and Developmental Disorders* 2010;40:988–999. [PubMed: 20130975]
- Biringen, Z.; Robinson, J.; Emde, R. *Emotional Availability (EA) Scales*. Unpublished Manual. 31998. www.emotionalavailability.com
- Chow S, Haltigan JD, Messinger DS. Dynamic infant–parent affect coupling during the Face-to-Face/Still-Face. *Emotion* 2010;10:101–114. [PubMed: 20141307]
- Gottman J, Levenson R. A valid measure for obtaining self report of affect. *Journal of Consulting and Clinical Psychology* 1985;53:151–160. [PubMed: 3998244]
- Granic I, Patterson G. Toward a comprehensive model of antisocial development. *Psychological Review* 2006;113:101–131. [PubMed: 16478303]
- Kerig, P.; Lindahl, K. *Family observational coding systems: Resources for systemic research*. Mahwah, NJ US: Lawrence Erlbaum Associates; 2001.
- Larrick RP, Soll JB. Intuitions about combining opinions: Mis-appreciation of the averaging principle. *Management Science* 2006;52:111–127.
- Lindahl K. Family process variables and children's disruptive behavior problems. *Journal of Family Psychology* 1998;12:420–436.
- Lindahl, KM.; Malik, NM. Unpublished manual. University of Miami; Miami, FL: 1996. *System for Coding Interactions and Family Functioning (SCIFF)*.
- Lindahl, K.; Malik, N. The System for Coding Interactions and Family Functioning. In: Kerig, P.; Lindahl, K., editors. *Family observational coding systems: Resources for systemic research*. Mahwah, NJ US: Lawrence Erlbaum Associates; 2001. p. 77-91.
- Messinger D, Cassel T, Acosta S, Ambadar Z, Cohn J. Infant smiling dynamics and perceived positive emotion. *Journal of Nonverbal Behavior* 2008;32:133–155. [PubMed: 19421336]
- Messinger DS, Mahoor M, Chow S, Cohn JF. Automated measurement of facial expression in infant–mother interaction: A pilot study. *Infancy* 2009;14:285–305. [PubMed: 19885384]
- NICHD Early Child Care Research Network. Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly* 1996;11:269–306.
- NICHD Early Child Care Research Network. Chronicity of maternal depressive symptoms, maternal sensitivity, and child functioning at 36 months. *Developmental Psychology* 1999;35:1297–1310. [PubMed: 10493655]
- Surowiecki, J. *The wisdom of crowds*. New York, NY: Doubleday; 2004.
- Waldinger RJ, Schulz MS, Hauser ST, Allen JP, Crowell JA. Reading others' emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. *Journal of Family Psychology* 2004;18:58–71. [PubMed: 14992610]