



HOKKAIDO UNIVERSITY

Title	An EXAMINATION OF PSYCHOMETRIC PROPERTIES AND VALIDITY OF THE TODDLER BEHAVIOR ASSESSMENT QUESTIONNAIRE
Author(s)	KUSANAGI, Emiko; 草薙, 恵美子; HOSHI, Nobuko et al.
Citation	乳幼児発達臨床センター年報, 19, 33-43
Issue Date	1997-03
Doc URL	https://hdl.handle.net/2115/25319
Type	departmental bulletin paper
File Information	19_P33-43.pdf



AN EXAMINATION OF PSYCHOMETRIC PROPERTIES AND VALIDITY OF THE TODDLER BEHAVIOR ASSESSMENT QUESTIONNAIRE

Emiko Kusanagi, Nobuko Hoshi, and Shing-Jen Chen
Hokkaido University

ABSTRACT

The Toddler Behavior Assessment Questionnaire (TBAQ) developed by Goldsmith was examined in terms of psychometric properties and validity: distribution of answer for each item, internal consistencies and discriminant properties of temperamental scales, and each convergence between maternal, paternal scales and laboratory measures. Although a few items were inadequate in distribution of answer, most items were found to be able to discriminate among Japanese children effectively. All scales proved to have high internal consistencies and acceptable discriminant properties. Concerning validity, there were low to moderate degree of agreement between mother and father report, and moderate convergence between maternal reports and laboratory measures only for Social Fear scale.

Key Words: temperament, questionnaire, validity, toddler.

INTRODUCTION

Since Alexander Thomas and Stella Chess initiated a longitudinal study of the implication of individual differences in temperament for psychological adaptation, the New York Longitudinal study (NYLS) (Thomas et al., 1963), there has been considerable interest among psychiatrists, developmental psychologists, and pediatricians in the study of temperament. More recently, researchers have begun to address the issue of influences of temperamental differences on the following areas; childhood social development of empathy, conscience and guilt (Eisenberg et al., 1996; Kochanska, 1991, 1993; Rothbart et al., 1994); adolescent personality traits (Caspi & Silva, 1995), behavior problems (Caspi, Henry, McGee, Moffitt, & Silva, 1995), and life course adjustment (Lerner et al., 1988).

Whilst temperament research has increased dramatically, a series of reviews have repeatedly pointed to theoretical and psychometric problems (Goldsmith & Rieser-Danner, 1990; Hubert et al., 1982; Slabach et al., 1991; Rothbart & Mauro,

The authors wish to thank Professor Hill Goldsmith and Ms. Kathy Lemery of the University of Wisconsin for help in this research. Correspondence concerning this article should be addressed to Emiko Kusanagi, Faculty of Education, Hokkaido University, Nishi 7, Kita 11, Kita Ku, Sapporo 060 Japan.

1990; Windle, 1988). Although the most popular form of temperament assessment in infancy and childhood is the parental report questionnaire, the typical questionnaire has been developed without sufficient attention to well established psychometric principles (Lerner, 1983). Some of these problems have been associated with the NYLS framework that is most widely used in temperament study. For instance, concerning the theoretical problems, Goldsmith and Campos (1982) have argued that the definition of temperament as "behavioral style" may apply to some of the NYLS dimension (e.g. Intensity), but that other dimensions are actually description of the content of behavior (e.g. Approach/Withdrawal). A series of reports of factor analysis of questionnaires that were developed by Carey and his associates according to the nine dimensional NYLS model, also call into question the theoretical assumptions underlying the NYLS approach. According to the result from a factor analysis of the Revised Infant Temperament Questionnaire (RITQ) with a sample of over 2400 Australian infants, although two dimensions emerged as relatively pure factors, the other factors were formed from various combinations or divisions of the NYLS dimensions (Sanson et al., 1987). Similarly in Japan, factor analysis of RITQ and Toddler Temperament scale (TTS) with a representative sample of 817 (RITQ) and 615 (TTS) children provided limited empirical support (Sugawara, 1994). The TTS is quite similar to the RITQ in format, with content changed slightly to make the items appropriate for the 12- to 36-months age range (Fullard et al., 1984). The Middle Childhood Temperament Questionnaire (MCTQ) for 8 to 12 year old children and the Behavior Style Questionnaire (BSQ) for 4 year old children also did not confirm the nine dimensions of NYLS (Buss & Plomin, 1984; McClowry et al., 1993). These consistent failure to duplicate the nine dimensional NYLS model suggests that it may be the theory rather than the instrument that is at fault (Slabach et al., 1991). As naturally expected, poor discriminant validity and low internal consistency of temperament scales occur more frequently in questionnaires developed by Carey and his associates (Goldsmith, Rieser-Danner, & Briggs, 1991; Slabach et al., 1991). Internal consistency estimates were below .70 for about half of the scales (Guerin et al., 1994; Fullard et al., 1984).

Failure to appropriately measure temperament constructs sometimes leads to an unfortunate need to reassess the validity of an entire body of literature using inappropriate instrument. If we could not find a relationship between a temperament dimension and future social development that is theoretically expected to relate each other, it would not be clear whether the fault lies with the false expectation or with the inappropriate measurements. Research in behavioral sciences often succeeds or fails depending on the quality of its assessment instruments (Goldsmith & Rothbart, 1991). Unfortunately, in Japan, the most widely used questionnaires are those developed by Carey and colleagues. We need to have psychometrically adequate measures of temperament.

Previously, we reported about the psychometric qualities of Rothbart's Children's Behavior Questionnaire (CBQ) (1988) that is developed for use with 3-7 years old children (Kusanagi, 1993b). As a result, the CBQ is found to have strong psychometric properties except a few scales. The companion instruments of the CBQ are Infant Behavior Questionnaire (IBQ), the Toddler Behavioral Assessment Questionnaire (TBAQ), and so on (Goldsmith & Rothbart, 1991). In this report, we examined the

psychometric properties and validity of the TBAQ constructed for 18-24-month old children by Goldsmith (1996).

Goldsmith & Campos presented their theoretical approach that defined infant temperament as the set of characteristic individual differences in the intensive and temporal response parameters of behavioral expression of affect-related states (Goldsmith & Campos, 1982). According to our results of temperament study using laboratory assessment, individual differences in positive emotion (joy/pleasure), anger, and fear are independent of each other (Kusanagi, 1993a; Hoshi et al., in press). The TBAQ is the only toddler questionnaire that captures positive emotionality independently of negative emotion and two negative emotionalities independently. The TBAQ measures temperamental dimensions of Activity Level, Tendency to Express Pleasure, Social Fearfulness, Anger Proneness, and Interest/Persistence. Utilizing 1,012 records, Goldsmith (1996) constructed the TBAQ by an iterative process of item generation intended to ensure content validity, by repeated item analysis focused in internal consistency and discriminant properties, and by scale revision. As a result, internal consistency reliability estimates typically exceeded .80 for each scale.

The psychometric qualities of the TBAQ were addressed in terms of three fundamental properties of scale and items in this report. First, the distribution of answers was examined for each item to confirm whether each item discriminated the subjects effectively. Second, coefficient alpha for each scale was computed in order to examine internal consistencies of the TBAQ scales. Finally, the intercorrelations among scales were computed to examine the discriminant properties of each scale. Interscale correlations are desired to be low. For the TBAQ, however, significant correlations were found among temperament scales.

The validity of the TBAQ were explored in terms of convergence between the maternal and paternal TBAQ scales, and convergence between the parental TBAQ scales and laboratory measures for three emotionalities. Thus far, for interparental agreement, moderate and highly variable relations appear to be the norm (Slabach et al., 1991). The degree of convergence between questionnaire and laboratory measurement was also reported to be moderate (Goldsmith & Rieser-Danner, 1990). In order to measure behavioral aspects of temperament, we used the locomotor version of Laboratory Temperament Assessment Battery (LAB-TAB) (Goldsmith & Rothbart, 1991). The locomotor version of LAB-TAB is designed by Goldsmith and Rothbart to make available a standardized instrument for laboratory assessment of temperament for 12- and 18-month-old without unusual or expensive equipments. Their first results suggested promising psychometric properties for the locomotor version of LAB-TAB (Goldsmith & Rothbart, 1991). The temperamental dimensions covered by LAB-TAB include Activity Level, Fearfulness, Anger Proneness, Interest/Persistence, and Joy-/Pleasure. In addition, we examined the agreement of laboratory measures with the mother's predictions about her child's behavior in laboratory settings to explore the ecological validity of LAB-TAB.

METHOD

Subjects :

Mothers whose children were about 18 month old who visited three health centers in Sapporo City for physical check-up were asked to complete the TBAQ. Some questionnaires were sent to mothers by mail from two health centers before physical check-up, and were collected when they visited the health centers. The rest of the questionnaires were handed to mothers at the third health center on visiting days, and were returned by mail. Return rate were 65.9% and 70.1% for the former, and 22.5% for the latter. A total of 483 mothers filled out the TBAQ. Among their children, 258 were boys and 222 were girls, and 3 children were not known in terms of sex. The age distribution of their children is presented in Table 1. There were three children whose age were not known. Approximately 20% of mothers of those that completed the TBAQ have agreed to participate in the laboratory study. Finally, we randomly

Table 1 Distribution of subject's age (N=480)

	Age				
	16 Months	17 Months	18 Months	19 Months	20 Months
Number of Subjects	2	84	378	14	2
%	0.4	17.4	78.4	2.9	0.4

selected 50 pairs of mothers and their children (23 boys and 27 girls) for laboratory assessment. Their average birth weight was 3129g, and all children were free of serious birth complications and congenital anomalies. Before the mother and her child left our laboratory, we gave the mother a self-addressed envelope and asked her to persuade the father to complete the questionnaire and send it back by mail. Thirty four fathers of these families have responded to the TBAQ.

The Toddler Behavior Assessment Questionnaire :

The TBAQ is comprised of 111 items that combine situations and responses theoretically postulated to be relevant to the target dimension. Several considerations were given to avoid the rater biases: (1) stating items clearly and simply; (2) minimizing the possible effects of social desirability by avoiding biasing language and using the strategy of asking the respondents to report the frequencies of discrete behaviors during a recent, specified time interval rather than making global judgements; and (3) disrupting possible response sets by reverse keying items. All items were directly translated into Japanese, except one item in which the name of a TV program ("Sesame Street") was deleted in the Japanese translation. Mothers or fathers were asked to indicate how often their children did the behavior described during the last month by circling one of the numbers. Numbers ranged from 1=never, through 4=about half the time, to 7=always, with X=does not apply (meaning, "I did not see my child in this situation"). Scale scores for the TBAQ were computed by summing all numerical item responses for a given scale. When a mother omitted an item, or checked the "does not apply" response option for an item, that item was not included.

The locomotor version of the Laboratory Temperament Assessment Battery :

The LAB-TAB is comprised of the 20 episodes, four per dimension, that are

used to assess temperament. We used a subset of the episodes to measure three emotionalities and Activity Level. We selected the following episodes; "Cognitive Assimilation Game (train)", "Modified Peek-a-boo Game", and "Reaction to Sound and Light Display (Simon)" episode for Joy/Pleasure; "Stranger Approach" and "Remote Controlled Toy" episodes for Fear; "Gentle Arm Restraint by Parent" and "Attractive Toy placed behind Barrier" episodes for Anger; and "Free Play" episode for Activity Level. The details of each episode procedure is described in the LAB-TAB.

According to Goldsmith & Rothbart (1991), we employed several practices to minimize carryover effects from one episode to another. We began with a nonstressful episode drawn from the Pleasure domains, and we avoid consecutive, potentially stressful episodes in the same room. We interspersed Fear and Anger episodes, and positioned the free play episode midway through the laboratory session. When the child became distressed, the next episode was resumed only after the child became soothed.

The laboratory sessions were videotaped, and rating were made from the replay by two raters. Five children were excluded from the analysis, as they became too distressed in more than four episodes. Following the guidelines for scoring the LAB-TAB, the longer episodes are typically divided in shorter intervals called epochs. Within each epoch or trial, a number of infant responses, such as smiling, reaching, or crying, are scored. Sometimes the presence or absence of a response is simply noted; however, more often parameters of the response, such as latency, duration, and intensity, are scored. The scoring is straightforward and subjective judgments were avoided as much as possible. Analysis for Activity Level of the LAB-TAB has not been performed yet, and was not reported here. Concerning the analysis of fear, in this report, we used data of only "Stranger Approach" episode, because the Fearfulness scale of the TBAQ is conceptualized as social fearfulness. "Reaction to Sound and Light Display (Simon)" episode was not utilized in the analysis, as about half of the children either did not express pleasure or displayed ambiguous expression. This episode seems to be inappropriate for eliciting Joy/Pleasure for Japanese children.

In terms of coding fear expression, when crying was observed, it was difficult to assess the expression using only fear coding system in the Affex (Izard et al., 1983). If his/her cry is interpreted as being due to fear, We assessed the expression using both coding system of distress and fear in the Affex, and adopted higher score of the two. As many children expressed intense distress at the epoch of being picked up and held in the "Stranger Approach" episode, we analyzed epochs before this for the peak of this measure. The parameter in which more than 80% of children had the same score was excluded from the analysis. The following are excluded parameters; The duration and latency of facial expression, and the duration and latency of bodily expression of fear ("Stranger Approach" episode); and the latency of bodily expression of anger ("Attractive Toy placed behind Barrier" episode).

All parameters were standardized before statistical analysis. Composite score of each episode was created by using the average of standardized scores except those which had not significantly positive relations with some of the other scores. Next, we

synthesized the higher order composite of each emotion by averaging composite score of episodes related to each emotion. This composite score of each emotion was used in the examination of the validity of the TBAQ.

To calculate interrater reliability, two well trained raters independently scored 10% of episodes that were selected randomly. The percentages of agreement between two raters were higher than 89% for all measures.

Following Goldsmith and Rothbart (1991), we administer a brief structured interview with the mother during warming up period prior to the LAB-TAB episodes. The

Table 2 Mean, Standard Deviation, Standard Error of the Mean, Actual Range of Scale Scores, and Cronbach's Alpha Estimates of Scores for Temperament Scales

Scales	Number	Mean	S. D.	S. E. M.	Actual range of scale scores	Alpha Estimates
AL	20	4.642	.669	.030	2.70-6.45	.85
PL	19	5.466	.687	.031	3.16-7.00	.92
SF	19	3.956	1.061	.048	1.08-7.00	.83
IN	22	4.076	.818	.037	1.22-6.21	.90
AP	28	4.060	.764	.035	1.92-6.21	.86

AL=Activity Level; PL=Pleasure; SF=Social Fear;
IN=Interest/Persistence; AP=Anger Proneness.

mother was explained about the episodes and was shown the pictures of the stimuli used at the episodes. Following this, we elicited her predictions about how her child would respond to the impending laboratory episodes on a 4-point scale (Definitely Yes, Probably Yes, Probably No, Definitely No). We calculated the average of rating scores for each episode, and then averaged them for each scale.

RESULTS

1. Distribution of responses for each item

The distribution of responses was checked out for each item. Seven items were marked "dose not apply" by more than 50% of the subjects. These items seem to be inappropriate for Japanese children. For example, at item that "When you were going out and your child did not want to stay with the regular sitter, how often did s/he pout or frown?", the percentage of "does not apply" was 80.1.

Nine items yielded same score by more than half of the mothers. For example, concerning the item that "When being tossed abut playfully or wrestled with, how often did your child smile?", 67% mothers replied that their children always smiled.

2. Internal consistencies

Table 3 Intercorrelations of TBAQ Scales (Pearson)

	AL	PL	SF	IN
PL	.18**			
SF	-.07	-.25**		
IN	-.22**	.23**	-.03	
AP	.36**	.14*	.25**	-.07

* $p < .01$, ** $p < .001$

The mean, standard deviation, standard error of the mean, range of scores, and a measure of internal consistency, coefficient alpha for each scale are shown in Table 2. The alpha ranged from .83 to .92, with a median of .86. The Pleasure scale evidenced very high internal consistency reliability (.92).

3. *Discriminant properties*

Intercorrelations across temperament scale scores are reported in Table 3. Seven of the ten interscale correlations were found to be significant. Activity Level scale was found to have moderate, positive correlation with Anger Proneness scale ($r=.36$), and modest, negative correlation with Interest scale. Pleasure scale showed low correlation with all the other scales. Among these, only correlation between Pleasure and Social Fear scale was negative. Two negative emotionarity scale (Anger Proneness and Social Fear) were positively interrelated. Social Fear scale did not correlate with either Interest or Activity Level scale. The relation of Interest scale with Anger Proneness scale was near zero.

4. *Correspondence of TBAQ maternal report with TBAQ paternal report and maternal interview*

Low to moderate correlations for parental agreement on the TBAQ were obtained (Table 4). There was moderate but significant agreement for Activity Level and Anger Proneness scale. Parental agreement for Interest/Persistence scale tended to be significant. It was evident that mother-father agreement for Pleasure and Social

Table 4 Correlations of TBAQ Maternal Report with TBAQ Paternal Report and Maternal Interview (Spearman)

Maternal	TBAQ Scale		Interview Maternal ^b
		Paternal ^a	
AL		.48**	.27 ⁺
PL		.21	.14
SF		.28	.25 ⁺
IN		.31 ⁺	—
AP		.46**	.20

** $p < .01$ ⁺ $p < .01$

^a N=34

^b N=50 for Activity, Pleasure & Anger r's, and 48 for Fear r's

Table 5 Correlations of Laboratory Measures with TBAQ Parental Report and Maternal Interview (Spearman)

Laboratory	TBAQ Scales		Interview Maternal ^c
	Maternal ^a	Paternal ^b	
PL	.18	— .07	.39**
SF	.33*	.23	.35**
AP	— .06	— .06	.27 ⁺

** $p < .01$, * $p < .05$, ⁺ $p < .01$

^a N=45 for Pleasure & Anger r's, and 40 for Fear r's

^b N=30 for Pleasure & Anger r's, and 27 for Fear r's

^c N=45

Fear scale was modest, but not significant.

There were modest correlations between maternal TBAQ report and interview. For Activity Level and Social Fear, agreement tended to approach significant level.

5. *Correspondence of laboratory measures with TBAQ parental report and maternal interview for the emotionality scales*

Correlation between maternal report on the TBAQ and the LAB-TAB measure was found to be significant only for Social Fear scale. The other correlations were very low or near zero. Correlations between laboratory measures and maternal interview concerning her child's reaction in the episode were low to moderate. For Pleasure and Social Fear scale, correlations was significant. For Anger Proneness scale, correlations tended to be significant.

DISCUSSION

The purpose of this study was to examine the TBAQ as developed by Goldsmith. He had made efforts of constructing scales that were theoretically grounded, conceptually independent, inclusive of varied facets of each temperament construct, internally consistent, and empirically distinctive. Concerning the distribution of answers, some items were found to be inappropriate for Japanese children. Many of them seemed to be due to the differences of child rearing practices between Japan and America. In Japan, when the mother goes out, it is rare that her child stay with the baby sitter. Although items that elicited highly skewed distributions of responses were eliminated in constructing the TBAQ, some items demonstrated such a distribution of responses in Japan.

According to a report by Goldsmith (1996), estimates of alpha were high for the scales of the TBAQ, ranging from .86 to .89 with the largest sample of 18-month-old. In this report, estimates of internal consistency attained high levels for all temperament scales. In general, measures of internal consistency are of moderate magnitude for temperament questionnaire (Hubert et al., 1982; Slabach et al., 1991). Estimate of internal consistency of the TTS used widely in Japan are low: alpha range = .56 to .71, median = .62 (Sugawara et al., 1994). Alpha estimates obtained in this study are of the highest magnitude among temperament questionnaires for toddlers.

Although temperament scales of the TBAQ are designated to be conceptually independent and empirically distinctive, Goldsmith and his associates take the view that interscale correlations should be low but not necessarily near zero (Goldsmith & Rothbart, 1991; Goldsmith, 1996). He noted that forcing zero intercorrelations is likely to make scale content unduly narrow. For example, they thought that Activity Level and Anger Proneness should be positively related simply because expression of anger often involves movement and arousal. Actually, this relation was found and was the strongest relation of all in the present study. The other correlations among the TBAQ scales were found to be similar to those reported by Goldsmith. There was an inverse relation between the Pleasure and the Social Fear scales. This seems to imply that broad hedonic tone may affect the responses as suggested by them. The expressions of pleasure and interest were positively correlated, and is interpreted by them as being due to the fact that expression of pleasure depends on engagement with the environment

and/or Interest/Persistence taps a milder aspect of a broad hedonically positive dimension. Pleasure also correlated positively with the Anger Proneness and Activity Level scales, for which high scores imply substantial behavioral activation. The evidence that the Social Fear scale was related with the Anger Proneness scale is inconsistent with our results in laboratory experiment (Kusanagi, 1993a; Hoshi et al., in press). Mothers may not be able to properly discriminate both negative emotions. Given that the correlations described above were low and explicable, the discriminant validity of the TBAQ seems to be acceptable.

One of the weakest aspects of parental rating is indexed by estimates of parental agreement (Martin & Halverson, 1991). The degree of parental agreement in this study fell within the range noted in the reviews of the temperament literature (Bate, 1987; Goldsmith & Rieser-Danner, 1990; Hubert et al., 1982; Slabach et al., 1991), and tended to be somewhat lower than those of the TBAQ reported by Goldsmith (1996). Some potential reasons why the degree of interparental agreement is low were suggested by researchers as follows: problems in adequacy of wording of scale items or instructions; rater response sets; differential weighting of rarely occurring child behaviors; differential amounts of contact leading to different ratings; different conceptual schemes used by different raters when evaluating specific child behaviors; the possibility that the child is displaying different temperaments for different individuals; differential influence of nontemperamental child characteristics to the temperament ratings of mothers versus fathers (Hubert et al., 1982; Slabach et al., 1991). As the level of agreement for scales that tapped active motoric behavior was found to be higher than the other scales, these more overt and noticeable aspects of temperament might elicit stronger consensus level among parents. It was demonstrated that the correlation between mothers and fathers on the IBQ Distress to Limitations (frustration/anger) scale was the highest of all scales (Goldsmith & Campos, 1990).

The evidence that there were significant, moderate relations between the laboratory affect measures and mother's interview-based predictions suggests that mothers are able to predict and rate their children's behaviors appropriately in a specific situation. Laboratory measures, however, did not evidence the relations with maternal TBAQ measures except the Social Fear scale. Maternal TBAQ scales showed only modest relation with the maternal interview measures. The first possible explanation of poor laboratory-questionnaire convergence is that the LAB-TAB settings may be different from situations where the mother was asked about the child's behavior in the TBAQ. Hagekull and his associates (1984) demonstrated that when observation were limited to situations specified by the parental rating instrument, validity coefficients were high. The second possibility is that strange laboratory rooms elicit too much tension for the child to express pleasure and/or anger as he/she might usually do. The third explanation is that reliabilities of laboratory measures may be questionable, as our laboratory measures of each dimension derived from one or two episodes. In home observation study, about six to eight observations are necessary to reliably measure the temperament dimensions (Sifer et al., 1994). It seems to be necessary to further examine the ecological validity and reliability of the LAB-TAB procedures for Japanese children.

To summarize, the TBAQ seems to have strong psychometric qualities, but yielded limited support of convergence with the laboratory measures.

REFERENCES

- Bates, J. E. (1987). Temperament in infancy. In J. D. Osofsky (Ed.), *Handbook of infant development* (2nd Ed., pp 555-578). New York: Wiley.
- Buss, A. H., & Plomin, R. (1984). *Temperament: Early developing personality traits*. Hillsdale, NJ: Erlbaum.
- Caspi, A., Henry, B., McGee, R. O., Moffitt, T. E., & Silva, P. A. (1995). Temperamental origins of child and adolescent behavior problems: From age three to age fifteen. *Child Development, 66*, 55-68.
- Caspi, A. & Silva, P. A. (1995). Temperamental qualities at age three predict personality traits in young adulthood: Longitudinal evidence from a birth cohort. *Child Development, 66*, 486-498.
- Eisenberg, N., Fabes, R. A., Murphy, B., Karbon, M., & Smith, M. (1996) The relations of children's dispositional empathy-related responding to their emotionality, regulation, and social functioning. *Developmental Psychology, 32*, 195-209.
- Fullard, D., McDevitt, S. C., & Carey, W. B. (1984). Assessing temperament in one-to-three-year old children. *Journal of Pediatric Psychology, 9*, 205-216.
- Goldsmith, H. H. (1996). Studying temperament via construction of the Toddler Behavior Assessment Questionnaire. *Child Development, 67*, 218-235.
- Goldsmith, H. H., & Campos, J. J. (1982). Toward a theory of infant temperament. In R. N. Emde & R. J. Harmon (Eds.), *the development of attachment and affiliative systems* (pp.161-193). New York: Plenum.
- Goldsmith, H. H., & Campos, J. J. (1990). The structure of temperamental fear and pleasure in infants: A psychometric perspective. *Child Development, 61*, 1944-1964.
- Goldsmith, H. H., & Rieser-Danner, L. (1990). Assessing early temperament. In C.R. Reynolds, & R. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context*. Vol. 2. New York: Guilford Press.
- Goldsmith, H. H., Rieser-Danner, L., & Briggs, S. (1991). Evaluating convergent and discriminant validity of temperament questionnaires for preschoolers, toddlers, and infants. *Developmental Psychology, 27*, 566-579.
- Goldsmith, H. H., & Rothbart, M. K. (1991). Contemporary instruments for assessing early temperament by questionnaire and in the laboratory. In J. Strelau & A. Angleitner (Eds.), *Explorations in temperament: International perspectives on theory and measurement* (pp.249-272). New York: Plenum Press.
- Guerin, D. W., & Gottfried A. W. (1994) Developmental stability and change in parent reports of temperament: A ten-year longitudinal investigation from infancy through preadolescence. *Merrill-Palmer Quarterly, 40*, 334-355.
- Hagekull, B., Bohlin, G., & Lindhagen, K. (1984). Validity of parental reports. *Infant Behavior and Development, 7*, 77-92.
- Hubert, N. C., Wachs, T. D., Peters-Martin, P., & Gandour, M. J. (1982). The study of early temperament: Measurement and conceptual issues. *Child Development, 53*, 571-600.
- Hoshi, N., Kusanagi, E., & Chen, S. J. (in press) An individual style does not exist for different emotional expressions in laboratory observations of infants. *Japanese Journal of Educational Psychology*. (in Japanese)
- Izard, C. E., Dougherty, L. M., & Hembree, E. A. (1983). *A system for identifying affect expressions by holistic judgments (Affex)*. Newark: University of Delaware, Computer Network Services and University Media Services.
- Kochanska, G. (1991). Socialization and temperament in the development of guilt and conscience.

- Child Development*, 62, 1379-1392.
- Kochanska, G. (1993). Toward a synthesis of parental socialization and child temperament in early development of conscience. *Child Development*, 64, 325-347.
- Kusanagi, E. (1993a) The structure of infant temperament: Emotional expression and the tendency to approach. *The Japanese Journal of Developmental Psychology*, 4, 42-50. (in Japanese)
- Kusanagi, E. (1993b) A psychometric examination of the Children's Behavior Questionnaire. *Annual Report of the Research and Clinical Center for Child Development*, 15, 25-33. Hokkaido University.
- Lerner, J. V., Hertzog, C., Hooker, K. A., Hassibi, M., & Thomas, A. (1988) A longitudinal study of negative emotional states and adjustment from early childhood through adolescence. *Child Development*, 59, 356-366.
- Lerner, J. V., & Lerner, R. M. (1983) Temperament and adaptation across life: Theoretical and empirical issues. In J. E. Bates, E. & O.G. Birns (Eds.), *Life-span development and behavior*, Vol. 5, 197-231.
- Martin, R. P., & Halverson, Jr. C. F. (1991). Mother-father agreement in temperament ratings: A preliminary investigation. In J. Strelau & A. Angleitner (Eds.), *Explorations in temperament: International perspectives on theory and measurement* (pp.235-248). New York: Plenum.
- McDevitt, S. C., & Carey, W. B. (1978). The measurement of temperament in 3-7 year old children. *Journal of Child Psychology and Psychiatry*, 19, 245-253.
- McClowry, S. G., Hegvik, R. L., & Teglasi, H. (1993). An examination of the construct validity of the middle childhood temperament questionnaire. *Merrill-Palmer Quarterly*, 39, 279-293.
- Rothbart, M.K. (1988). Children's Behavior Questionnaire. Unpublished manuscript, University of Oregon.
- Rothbart, M. K., Ahadi, S. A., & Hershey, K. L. (1994) Temperament and social behavior in childhood. *Merrill-Palmer Quarterly*, 40, 21-39.
- Rothbart, M. K., & Mauro, J. K. (1990). Questionnaire approach to the study of infant temperament. In J. W. Fagen & J.Colombo (Eds.), *Individual Differences in infancy: Reliability, stability and prediction* (pp. 411-430). Hillsdale, NJ: Erlbaum.
- Sanson, A. V., Prior, M., Garino, E., Oberklaid, F., & Sewell, J. (1987). The structure of infant temperament: Factor analysis of the revised infant temperament questionnaire. *Infant Behavior and Development*, 10, 97-104.
- Sanson, A. V., Smart, D. F., Prior, M., Oberklaid, F., & Pedlow, R. (1994) The structure of temperament from age 3 to 7 years: Age, sex, and sociodemographic influences. *Merrill-Palmer Quarterly*, 40, 233-252.
- Seifer, R., Sameroff, A. J., Barrett, L. C., & Krafchuk, E. (1994) Infant temperament measured by multiple observations and mother report. *Child Development*, 65, 1478-1490.
- Slabach, E. H., Morrow, J., & Wachs, T.D. (1991) Questionnaire measurement of infant and child temperament: Current status and future directions. In J. Strelau & A. Angleitner (Eds.), *Explorations in temperament: International perspectives on theory and measurement* (pp.205-234). New York: Plenum.
- Sugawara, M., Shima, S., Toda, M., Sato, T., & Kitamura, T., (1994). Behavioral characteristics in early childhood: An investigation of RITQ and TTS (Japanese Version). *Japanese Journal of Educational Psychology*, 42, 315-323. (in Japanese)
- Thomas, A., Chess, S., Birch, H., Hertzog, M., & Korn, S. (1963). *Behavioral individuality in early childhood*. New York: New York University Press.
- Windle, M. (1988) Psychometric strategies of measures of temperament: A methodological critique. *International Journal of Behavioral Development*, 11, 171-201.