Studying Temperament via Construction of the Toddler Behavior Assessment Questionnaire

H. H. Goldsmith

University of Wisconsin—Madison

GOLDSMITH, H. H. Studying Temperament via Construction of the Toddler Behavior Assessment Questionnaire. CHILD DEVELOPMENT, 1996, 67, 218–235. The Toddler Behavior Assessment Questionnaire (TBAQ) was constructed by an iterative process of item generation intended to ensure content validity, by repeated item analyses focused on internal consistency and discriminant properties, and by scale revision. During the construction and initial validation processes reported in this article, data from 1,012 records were utilized. Internal consistency reliability estimates typically exceeded .80 for each scale. Evidence for convergent validity with other temperament questionnaires and for longitudinal stability was also obtained. Besides yielding a promising instrument, this assessment research has conceptual ramifications. For instance, components of negative affectivity (anger proneness and fearfulness) were independent, and item analyses suggested that shyness and other fears were independent as well. Consistent with most current views of temperament, the TBAQ temperament scales revealed some relationship and/or contextual specificity, as exemplified by the finding of only moderate parental agreement. The rank ordering on most temperament dimensions was impressively preserved from age 12 months, when the Infant Behavior Questionnaire (IBQ) was used, to age 18 months, when the TBAQ was used (especially when subtle differences between the IBQ and TBAQ were taken into account). Most of the analyses benefit from replication.

The best way to construct a psychometric instrument is slowly. Years of careful conceptualization, design, testing, revision, and validation should yield not only a useful instrument but also some new insights into the domain being assessed. This is hardly a new idea (Jessor & Hammond, 1957; Loevinger, 1957), and it is true whether the instrument is a questionnaire, an interview schedule, a behavioral test, or a physiological assessment; it is true whether the domain is perceptual, cognitive, emotional, or social. In the domain of adult personality, Tellegen and Waller (in press) have documented the fruitfulness of viewing test construction in this way, as they elucidated the self-report domain of adult personality via construction of the Multidimensional Personality Questionnaire. I pursued a similar approach in the parental report domain via construction and initial validation of the Toddler Behavior Assessment Questionnaire (TBAQ).

The TBAQ’s domain is early temperament, a field of research that has been preoccupied with definitional and measurement issues, neither of which this article treats extensively. Fortunately, several excellent treatments of these issues are available. Buss and Plomin, Goldsmith, Rothbart, and

The research summarized in this article owes much to a series of faculty, graduate student, and research assistant collaborators at the Universities of Denver (1980–1981), Texas at Austin (1881–1984), Oregon (1985–1992), and Wisconsin—Madison (1992–1993). Mary K. Rothbart’s Infant Behavior Questionnaire served as the model for the TBAQ. Joseph P. Campos sponsored Goldsmith as a postdoctoral fellow during the initial work. His advice was instrumental in the approach to temperament assessment represented by the TBAQ and associated instruments. Donna L. Bradshaw and Loretta A. Rieser-Danner participated while graduate students at the Universities of Denver and Texas at Austin, respectively. They contributed ideas and validity data to the project. Research assistants who worked directly on the project and whose ideas were directly translated into items were Patricia East, Kate Duncan, Karen Jaco, and Teresa Elliott. Mark Layman and Teresa Elliott executed some of the initial statistical analyses. A number of other research assistants and graduate students contributed in various ways, Lori Bowden, Jennifer Alansky Mauro, Nancy Benson Fischette, and Kathryn Lemery being foremost among them. Jennifer Urbano of Vanderbilt University offered an important data analytic suggestion. Michael Crowley of the University of Oregon collected the data in Sample 10 as part of a more extensive graduate research project. Support was provided by the Foundation for Child Development, the Spencer Foundation, the MacArthur Foundation, NICHD, and NIMH.

[Child Development, 1996, 67, 218–235. © 1996 by the Society for Research in Child Development, Inc. All rights reserved. 0009-3920/96/6701-001/00.00]
Thomas and Chess contrasted their four theories of temperament in roundtable discussion and subjected them to critique and synthesis by Hinde and McCall (Goldsmith et al., 1987). Two recent handbooks of temperament research present a variety of theoretical approaches, both to the overarching domain of temperament and to key constructs under the temperament rubric, such as difficulty and inhibition (Kohnstamm, Bates, & Rothbart, 1989; Strelau & Angleitner, 1991).

At the risk of oversimplification, contemporary temperament theories can be categorized as emphasizing the energetic aspects of behavior, including regulatory functions (Rothbart, 1989; Strelau, 1983); the stylistic aspects of behavior (Thomas & Chess, 1977); the early-appearing, heritable aspects of personality (Buss & Plomin, 1984); or the emotional aspects of behavior (Goldsmith & Campos, 1982, 1986; Merahbian, 1991). These theoretical approaches have been compared and critiqued extensively (Bates, 1987; Goldsmith & Campos, 1982; Seifer & Sameroff, 1986; Strelau, 1987). Obviously, they are not mutually exclusive. For example, emotional traits can also appear early in development, be heritable, involve energy expenditure, have regulatory functions, and contribute to stylistic distinctiveness of individuals.

The enterprise of assessing temperament incorporates parent, teacher, and self-report questionnaires, as well as interviews, observations, and laboratory-based procedures, for eliciting temperamental reactions. The laboratory approach is enjoying some prominence, with batteries of varying scope available (Goldsmith, Reilly, Lemery, Longley, & Prescott, 1993; Goldsmith & Rothbart, 1991; Kagan, Reznick, Clarke, Snidman, & García-Coll, 1984; Matheny, 1991; Seifer & Sameroff, 1986). Although the objectivity, clear behavioral referents, and standardization of laboratory-based approaches are attractive, their expense and possible artificiality are not. There remains an important role for caregiver-report measures of early temperament, as recognized in several reviews of temperament assessment (Angleitner & Riemann, 1991; Bates, 1989; Campos, Barrett, Lamb, Goldsmith, & Stenberg, 1983; Goldsmith & Rieser-Danner, 1990; McDevitt, 1986; Rothbart & Mauro, 1990; Slabach, Morrow, & Wachs, 1991).

Within this literature, the pros and cons of using parents as informants have been discussed at length (e.g., Bates & Bayles, 1984; Rothbart & Goldsmith, 1985; Seifer & Sameroff, 1986). The advantages include the parent’s unparalleled knowledge of the child’s behavior, the ability to inquire about many domains of temperament in a short time, and low cost. The difficulties of using parents as informants are that they vary (1) in how well they comprehend questions and use rating scales, (2) in their psychological state while completing the questionnaire, (3) in their recall of the child’s behavior, (4) in their acquaintance with other children who might provide a basis of comparison, and (5) in the extent of their actual knowledge of the child’s behavior. It is also widely recognized that parental views of temperament are probably conditioned by their own styles of interacting with the child. Finally, parents might be biased by the same response styles that have been investigated in the realm of adult self-report of personality, including tendencies to respond in a socially desirable fashion, tendencies to agree with items (acquiescence), and even “cries for help” when the investigator is perceived as a potential source of counseling for a child with problems. Many of these potential difficulties can be negotiated by careful item writing (Foddy, 1993).

Narrowing our focus to temperament assessed via questionnaire during the toddler period, there are two general approaches to item writing. One approach uses rather general items (“child is always on the go”; “child is afraid of novel objects”) in questionnaires that can apply to toddlers as well as infants and older children (Buss & Plomin, 1984; Rowe & Plomin, 1977; Windle & Lerner, 1986). Another approach is to use items specifically tailored for the toddler period. This approach is represented in the TBAQ as well as other questionnaires (Bates, Freeland, & Lounsbury, 1979; Fuldard, McDevitt, & Carey, 1984; Hagekull, 1985). There are obviously trade-offs in the choice between these approaches that must be evaluated according to the goals of the research.

The TBAQ measures temperamental dimensions of Activity Level, tendency to express Pleasure, Social Fearfulness, Anger Proneness, and Interest/Persistence. I initially designed the TBAQ for use with 18–24-month-old children but have found it useful for an age range of 16–36 months. The upper range has been extended by several months in some studies.

The purpose of this article is to describe the construction of the TBAQ, to provide ini-
tial evidence of its reliability and validity, and, from this methodological work, to highlight some substantive findings.

Method

Sample

This article reports data from 11 samples (sometimes including more than one respondent and sometimes more than one occasion). The samples and their use are previewed in Table 1. Each of the samples is a broad working/middle socioeconomic class, mixed-gender sample, with the proportion of Caucasians ranging from approximately 85% to 95% and minimal representations of Hispanics, Asian-Americans, and African-Americans. These subjects were ascertained from birth records and recruited by telephone or letter, with the mother as respondent, unless otherwise specified.

Families in Sample 1 were recruited from a psychology department infant subject pool, which was based on a search of state birth records. Sample 2 comprised 204 protocols from parents who had earlier participated in another study (see Goldsmith & Campos, 1990, for a report of this earlier study). Over 75% of the Sample 2 records were from twins, and somewhat more mothers than fathers participated.

Samples 3, 4, and 5 were from the same study, done in cooperation with a regional organization of mothers-of-twins clubs. For evaluating the TBAQ, I divided this group into three samples: (1) Sample 3, comprising one twin from each pair; (2) Sample 4, a replicate sample consisting of the other twin from each pair; and (3) Sample 5, the singleton, toddler-age siblings of other twins outside the TBAQ age range. Note that Samples 3 and 4 are replicates in more than the usual sense because the same mothers are respondents in both samples. The representativeness of twins is an issue in studies such as this. Because mothers-of-twins clubs recruit new members extensively, they are thought to be fairly representative of volunteer samples when the twins are young (when their twins grow older, many members leave the clubs). Prior work suggests that twins do not differ from singleton individuals on caregiver-report temperament when assessed with the similar IBQ (Goldsmith & Campos, 1990).

Samples 6, 7, 8, 9, 10, and 11 were recruited from birth announcements in a local

<table>
<thead>
<tr>
<th>Sample No. in Text</th>
<th>No. of TBAQ Records Used</th>
<th>Age (Months)</th>
<th>Summary of How the Sample Was Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99</td>
<td>18–24</td>
<td>Initial item analysis of early three-scale version</td>
</tr>
<tr>
<td>2</td>
<td>204</td>
<td>21–22</td>
<td>Item analyses; relation to IBQ given earlier; parental correspondence</td>
</tr>
<tr>
<td>3</td>
<td>93</td>
<td>16–36</td>
<td>Item analysis</td>
</tr>
<tr>
<td>4</td>
<td>93</td>
<td>16–36</td>
<td>Item analysis</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>16–36</td>
<td>Item analysis</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>18–24</td>
<td>Item analysis; standardization; discriminant validity</td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>18–24</td>
<td>Item analysis; standardization; discriminant validity; relation to ICQ</td>
</tr>
<tr>
<td>8a</td>
<td>66</td>
<td>18</td>
<td>Discriminant validity; relation to social-emotion scales</td>
</tr>
<tr>
<td>8b</td>
<td>37</td>
<td>48</td>
<td>Subgroup of 8a followed up at 4 years for stability; relation to social-emotion scales</td>
</tr>
<tr>
<td>9a</td>
<td>57 mothers</td>
<td>18–28</td>
<td>Parental correspondence</td>
</tr>
<tr>
<td>9b</td>
<td>57 fathers</td>
<td>18–28</td>
<td>Relation of IBQ and TBAQ; discriminant validity</td>
</tr>
<tr>
<td>10</td>
<td>81</td>
<td>14–21*</td>
<td>Standardization; discriminant validity; continuity from IBQ to TBAQ (51/78 were tested at 12 months on the IBQ)</td>
</tr>
<tr>
<td>11a</td>
<td>78 mothers</td>
<td>18</td>
<td>(Subgroup of 11a) Parental correspondence</td>
</tr>
<tr>
<td>11b</td>
<td>27 fathers</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

*Note.—Total number of TBAQ records used: 1,012.

*66% of the subjects in this sample were 15, 16, or 17 months of age.
newspaper. These six samples participated in studies with various goals, the common denominator being inclusion of the TBAQ. Only the TBAQ results and other temperament questionnaire results are included in this report. Table 1 shows the age composition of each sample and the sample’s purpose. Other relevant features of the samples are mentioned below: Sample 6 included only toddlers with a sibling between ages 3 months and 6 years. Sample 7 was composed almost entirely of families with only one child. Mothers in Sample 8 completed the TBAQ at both ages 18 months and 4 years. Sample 9 was more highly selected than the others because substantial cooperation of fathers was required (see Crowley, Goldsmith, & McNeil, 1991). Sample 10 was used only for the questionnaire-based research reported in this article; the children were in the transitional age range between the IBQ and TBAQ, and mothers completed both questionnaires at one time. The Sample 11 subjects cooperated in other laboratory-based procedures; with a few exceptions, mothers completed the temperament questionnaires before visiting the laboratory.

Overview of Procedures

The methods used in construction and validation are a major topic of this article, and thus they are covered below. In broad outline, scales and items were generated rationally according to preconceived notions about important domains of temperament and their constituent behaviors. I revised these notions as the empirical data dictated. The scales were derived via item analysis in a classic manner well described by Wiggins (1989), for example. This process maximized internal consistency and limited nonorthogonality. The process was iterative, as might be expected from the multiple samples described above. An alternative to item analysis would have been exploratory factor analysis of items (which would have maximized orthogonality), but this was rejected for three reasons: (1) requirements of subject: variable ratio are more stringent for factor analysis, and thus fewer rounds of replication and revision would have been possible; (2) identification of the structure of temperament from some corpus of items was not my goal; and (3) factor analysis of items to form scales generally leads to factors (scales) of decreasing reliability—even after rotation—and a chief goal was for the TBAQ scales to have equivalent psychometric characteristics.

The Iterative Construction Process, Content Validity, and the Nature of Temperament

Theoretical Approach

The evolving theory that guided the TBAQ’s construction regards temperament as the individual differences aspect of emotionality in early development (Goldsmith, 1993; Goldsmith & Campos, 1982, 1986). Thus, all the TBAQ scales relate overtly to emotional expression, with the apparent exception of Activity Level. This exception may be more apparent than real, however. Goldsmith and Campos (1982) hypothesized that the child’s characteristic activity level may be related to a general level of emotional arousal. Even if this hypothesis ultimately fails to find support, a measure of activity is useful in most temperament or emotion research programs. Theoretically, approach and withdrawal tendencies are implicated in many aspects of emotionality and temperament (e.g., Davidson, 1992; Fox & Davidson, 1984), with approach tendencies often involving motoric activation and withdrawal tendencies often involving decreased motoric activity or, in extreme cases, even freezing. Operationally, the responses to many items on other scales involve motoric activity in one way or another. Thus, it may prove informative to control statistically for the effects of activity-level differences when correlating another TBAQ scale to some external criterion.

Some readers might be surprised to see interest/persistence designated as affect related; however, interest satisfies most criteria for a “basic” emotion (Ekman, 1992), including a specific facial expression that is one cornerstone of discrete affect theory (Izard, 1993) and a particular appreciation of organism-environment relationship that is one cornerstone of functionalist approaches to emotion (Barrett & Campos, 1987; Lazarus, 1991). Cognitive developmentalists also view interest as affect—an important affect that drives cognitive development. For instance, Piaget, following Claparède (1909, cited in Piaget, 1981), viewed interest as an anticipatory regulation of cognitive engagement—a “dynamogenisateur” of action. Persistence, as operationalized in the TBAQ, is the individual differences counterpart of interest; that is, stable and consistent patterns of interest define persistence.

The TBAQ scales are not necessarily hypothesized to reflect the most “basic” as-
pects of temperament, whether "basic" is defined in a biological or factor-analytic sense. Rather, they are designed to provide an assessment of individuality in the emotional domain, which is the focus of extensive research (Ekman & Davidson, 1994; Fox, 1994; Lazarus, 1991; Lewis & Haviland, 1993; Oakley, 1993; Saarni & Harris, 1989; Stein, Leventhal, & Trabasso, 1990; Thompson, 1990). The focus on affects such as anger, fear, pleasure, and interest does not preclude alternative concepts such as approach-avoidance, hedonic tone, arousal level, or emotion regulation.

The Construction Process and Content Validity Considerations

The TBAQ was initially developed as an instrument to follow up subjects tested in infancy with Rothbart's (1981) Infant Behavior Questionnaire (IBQ); however, it has evolved into a general purpose temperament inventory. The construction technique was a three-part, iterated process. First, item-writing teams constructed items that combined situations and responses theoretically postulated to be relevant to the target dimension. The team solicited opinions from mothers and other researchers during this period of item writing. Second, the set of candidate items was administered to samples of mothers with children in the appropriate age range. I computed item statistics and eliminated items that showed low correlations (generally < .25) with their intended scale or that correlated relatively highly (generally > .15) with another scale. Items which, when dropped from the scale, resulted in increased scale alpha estimates were also eliminated. After considering the content of omitted items, the item-writing team constructed new items and administered the revised scales to a new sample. This process continued for a total of four waves of data collection, with a total of seven samples. However, I used the Interest/Persistence and Anger Proneness scales in only the last two waves, which comprised a total of five samples. Several considerations were balanced during item writing and item analysis: (1) striving for variety of context and response while adhering to theoretical considerations; (2) using only situations and responses that would apply throughout the toddler period and that would not differ by gender; (3) stating items clearly and simply—the Flesch reading level of the entire final item set is grade 8.0; (4) minimizing the possible effects of social desirability by avoiding biasing language and using Rothbart's (1981) strategy of asking the respondents to report frequencies of discrete behaviors during a recent, specified time interval rather than make global judgments; (5) disrupting possible response sets by reverse-keying items (31% were reverse-keyed in the final version); (6) eliminating items that elicited highly skewed distributions of responses; and (7) striving for a distribution of item difficulties on each scale that was peaked in the middle of the 7-point rating scales, with some representation of high and low difficulty items (as reflected in item means) to allow better differentiation at the temperament extremes. The goal was to achieve final versions of the scales with about 20 items each and to avoid undue narrowing of content while retaining high internal consistency.

Throughout the process, I examined the implications of omitted items for the basic conceptualization of each dimension. For instance, items concerning fearful reactions to spiders and insects, masks, the dark, thunderstorms, other loud noises, strange dogs, and other animals repeatedly failed to correlate with more socially oriented fear items. Therefore, the Fear scale was renamed Social Fearfulness. Thus, this scale measures shyness rather than wariness novel objects or inhibition in nonsocial situations. In this regard, the TBAQ Social Fearfulness scale is not comparable to IBQ Distress to Novelty scale (sometimes referred to as a fear scale); the IBQ Distress to Novelty scale contains only five items related to social situations. A direction worth pursuing in future research is whether items reflecting responses in nonsocial incentive conditions cohere within the toddler age range.

There are other important content considerations. The Interest/Persistence scale primarily measures sustained attention to an ongoing activity rather than responses to distractions. Thus, it does not cover the entire domain of attention-related characteristics. Also, the scale does not tap the connotation of the term "persistent" that means "stub-
born." Another content consideration relates to an observation that the Anger Proneness and Activity Level scales correlated more strongly than desirable in Samples 3, 4, and 5. This led the item-writing team to design new activity items with more positive hedonic connotations. Other similar adjustments in wording were made to "balance" scale content.  

Description of Scale Content

The scale content can best be appreciated by studying the entire item pool; however, the extended scale definitions and sample items below may be useful.

Activity Level: Limb, trunk, or locomotor movement during a variety of daily situations, including free play, confinement, or quiet activities. ("When playing on a movable toy, how often did your child attempt to go as fast as s/he could?")

Pleasure: Smiling, laughter, and other hedonically positive vocalizations or playful activity in a variety of nonthreatening or familiar situations. ("When in the bathtub, how often did your child babble or talk happily?")

Social Fearfulness: Inhibition, distress, withdrawal (vs. approach), or signs of shyness in novel or uncertainty-provoking situations of a social nature. ("When your child was being approached by an unfamiliar adult while shopping or out walking, how often did your child show distress or cry?")

Anger Proneness: Crying, protesting, hitting, pouting, or other signs of anger in situations involving conflict with another child or the caregiver. ("When you did not allow your child to do something for her/himself, for example, dressing or getting into the car seat, how often did your child try to push you away?")

Interest/Persistence: Duration of task engagement in ongoing solitary play or other activities. ("How often did your child play alone with his/her favorite toy for 30 minutes or longer?")

Response to the items is indicated on a seven-point scale ranging from (1) "Never" to (4) "About half the time" to (7) "Always," with a provision for "does not apply." Higher scores always reflect higher standing on the trait named by the scale.

Item Analyses: Internal Consistency and Discriminant Properties

Internal consistency.—Throughout the various phases of construction, estimates of internal consistency (alpha) have been high compared to other instruments of this genre. Table 2 provides replicated alpha estimates, derived by simply averaging the alphas from two mixed-age samples (6 and 7) and the largest sample of 18-month-olds (Sample 11a), as well as other psychometric information. High internal consistency estimates carry important theoretical ramifications for temperament theory. They imply coherence among behaviors specified as temperament-related by the test constructor; such coherence is a defining feature of temperament. Of course, high internal consistency could be achieved by using tautologous items (not the case in the TBAQ) or by using many less highly correlated items (which is the case).

Table 2 also shows the average item-scale correlations (with all items included in the scale score), which can be viewed as internal consistency estimates independent of scale length. This statistic is useful for comparison with scales from other questionnaires.

Another important feature of the TBAQ is that scales are approximately equal in length, internal consistency (alpha), and level of item-scale interrelation, as shown in Table 2. These properties are desirable because differential correlations with other measures are unlikely to be due to differential scale reliability, a problem with some other temperament questionnaires. It is also important to note that some internal consistency estimates in the literature are based on derivation samples, so they might be artifactualy high, and this is not the case for the Table 2 statistics.

Discriminant properties.—Table 3 shows average interscale correlations derived from five samples. The upper section of Table 3 summarizes data from samples ranging in age, whereas the lower section summarizes data from samples with 18-month-old toddlers. There were minimal differences among the samples averaged in Table 3 and also few differences depending on age ranges.

As Table 3 shows, the TBAQ scales are largely independent, thus supporting the TBAQ's discriminant validity. Taking into account effect sizes and the pattern of replication, the three bold-faced correlations are notable. Activity and Anger tend to corre-

---

2 A more detailed description of the scale construction process and the actual questionnaire with scoring information are available from the author.

3 One implication of this is that interpretation of factor analyses of the TBAQ scales should be cautious. The effort to remove common content variance in the scales suggests that some covariation might arise from noncontent sources.
late positively, undoubtedly because angry behavior often involves movement and arousal, in contrast to fearfulness. The expressions of pleasure and interest are also positively correlated, perhaps because expression of pleasure depends on engagement with the environment and/or because Interest/Persistence taps a milder aspect of a broad hedonically positive dimension. Social Fearfulness is negatively correlated with Pleasure, lending some support to an approach/avoidance interpretation of this source of covariation.

In summary, this “monomethod” (Campbell & Fiske, 1959) version of discriminant validity of the TBAQ is acceptable, given that the interscale correlations are explicable according to other conceptualizations of emerging personality (i.e., the importance of overarching dimensions of positive and negative affectivity) and the role of action (captured in the Activity Level scores) in expressing strong emotion. Thus, I do not subscribe to the view that temperament scales should necessarily have zero intercorrelations. Forcing zero intercorrelations is likely to make scale content unduly narrow. And, of course, such independence may not obtain in nature.

This series of studies, then, resulted in a questionnaire of acceptable psychometric qualities and also provided several insights into the nature of temperament. Although a systematic construction process is essential, it does not show that the TBAQ is useful. Assessing the validity of any instrument is a continuing process. Below, I report initial validity studies wherein assessment is confined to the questionnaire domain.

### Table 2

<table>
<thead>
<tr>
<th>Scale</th>
<th>No. of Items</th>
<th>Alpha (Samples 6 and 7)</th>
<th>Alpha (Samples 11)</th>
<th>Average Item-Totai r (Samples 6 and 7)</th>
<th>Average Item-Totai r (Samples 11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL—Activity Level</td>
<td>20</td>
<td>.78</td>
<td>.86</td>
<td>.44</td>
<td>.53</td>
</tr>
<tr>
<td>PL—Pleasure</td>
<td>19</td>
<td>.80</td>
<td>.86</td>
<td>.47</td>
<td>.53</td>
</tr>
<tr>
<td>SF—Social Fearfulness</td>
<td>19</td>
<td>.83</td>
<td>.87</td>
<td>.49</td>
<td>.55</td>
</tr>
<tr>
<td>AN—Anger Proneness</td>
<td>28</td>
<td>.82</td>
<td>.88</td>
<td>.41</td>
<td>.48</td>
</tr>
<tr>
<td>IN—Interest/Persistence</td>
<td>22</td>
<td>.79</td>
<td>.89</td>
<td>.45</td>
<td>.54</td>
</tr>
</tbody>
</table>

*See text for explanations.

b Data from Samples 6 and 7 averaged.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Activity Level</th>
<th>Pleasure</th>
<th>Social Fearfulness</th>
<th>Anger Proneness</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Average values from mixed-age samples: b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity Level</td>
<td></td>
<td>.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pleasure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Fearfulness</td>
<td></td>
<td>-.10</td>
<td>-.24</td>
<td></td>
</tr>
<tr>
<td>Anger Proneness</td>
<td></td>
<td>.27</td>
<td>.07</td>
<td>-.01</td>
</tr>
<tr>
<td>Interest/Persistence</td>
<td></td>
<td>-.03</td>
<td>.37</td>
<td>-.07</td>
</tr>
<tr>
<td>II. Average values from 18-month-old samples: c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity Level</td>
<td></td>
<td>.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pleasure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Fearfulness</td>
<td></td>
<td>.15</td>
<td>-.21</td>
<td></td>
</tr>
<tr>
<td>Anger Proneness</td>
<td></td>
<td>.48</td>
<td>.17</td>
<td>.13</td>
</tr>
<tr>
<td>Interest/Persistence</td>
<td></td>
<td>-.23</td>
<td>.26</td>
<td>-.14</td>
</tr>
</tbody>
</table>

a Averaged correlations with underscore show a replicated pattern that is discussed in the text.
b Samples 6, 7, 9, and 10; total N = 237.
c Samples 8a and 11a; total N = 144.
Initial Validity Studies

Construct validation, viewed as including concurrent and predictive validity as well as testing of theoretical propositions, is a continuing process. The initial validity studies reported in this paper are confined to temperament questionnaires.

Relations with Other Temperament Measures

Correlation with “difficulty.”—The most widely known—and somewhat controversial—construct in the temperament literature in this country is “difficulty” (Bates, 1980; Thomas & Chess, 1977). The centrality of difficulty in the clinical literature demands that any new measure of temperament be related to it. I chose Bates’s Infant Characteristics Questionnaire (ICQ) (Bates et al., 1979) as a well-validated multifactor measure of difficulty. Of course, difficulty as measured by the ICQ is probably somewhat different than difficulty measured by other scales, such as the Toddler Temperament Scale (TTS) (Fullard, McDevitt, & Carey, 1984). In Sample 7, I administered both the TBAQ and the ICQ. The ICQ has four factor-analytically derived subscales at age 13 months: the largest factor, Fussy/Difficult; and three other factors, Persistence, Unadaptability, and Unsoiability. An overall difficulty index can also be calculated from the ICQ. There is substantial continuity of the main difficulty factor across the ages from 6 to 13 to 24 months (Bates, 1992). The upper portion of Table 4 shows the TBAQ-ICQ intercorrelations. Three TBAQ scales showed significant correlations with the ICQ overall difficulty index: Activity Level (.57), Anger Proneness (.54), and Pleasure (.22). Unsurprisingly, these same three scales were similarly correlated with the large first factor of the ICQ, Fussy/Difficult. The most notable other correlations were TBAQ Activity Level and Anger with ICQ Persistence. Each of these correlations was recomputed with correction for age and sex with minimal change.

The overall pattern of results suggests that maternal perception of difficulty is largely a reflection of positive versus negative affect expression in the toddler (as perceived by the mothers, of course). The negative affect involved in this relationship is anger rather than fear; thus, anger expres-

<table>
<thead>
<tr>
<th>TABLE 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORRELATIONAL ANALYSES RELATING THE TBAQ SCALES TO OTHER QUESTIONNAIRES</td>
</tr>
<tr>
<td>I. RELATION TO BATES’S ICQ MEASURES OF DIFFICULTY*</td>
</tr>
<tr>
<td>TBAQ SCALES</td>
</tr>
<tr>
<td>Overall Difficulty</td>
</tr>
<tr>
<td>Fussy/Difficult subscale</td>
</tr>
<tr>
<td>Unadaptability subscale</td>
</tr>
<tr>
<td>Persistence subscale</td>
</tr>
<tr>
<td>Unsoiability subscale</td>
</tr>
</tbody>
</table>

II. RELATION TO ROTHBART, AHADI, & HERSHEY’S ‘SOCIALIZATION-RELEVANT TRAITS’b

| TBAQ SCALES |
| SOCIALIZATION-RELEVANT TRAIT SCALES |
| Activity Level | Pleasure | Social Fearfulness | Anger Proneness | Interest/Persistence |
| 18m | 4y | 18m | 4y | 18m | 4y | 18m | 4y |
| Aggression | .23 | .38* | .07 | -.29* | -.10 | .24* | .13 | .58* | -.06 | -.16 |
| Empathy | -.26 | -.19 | .17 | .32* | .01 | -.17 | -.29* | .02 | .11 |
| Guilt/Shame | -.32* | -.26* | .16 | .19 | .30* | .53* | -.10 | -.19 | -.01 | .11 |
| Help-seeking | -.13 | .04 | .08 | .09 | .14 | .25* | .02 | -.01 | -.34* | -.19 |

* Sample 7, N = 54 for data on both questionnaires.

b Sample 8, N = 37 for 18-month correlations and N = 49 for 4-year-olds. Correlations underscored in the lower section of the table were predicted from item content similarity or from previous findings (see text).

*p < .05.
sion appears to be the emotional core of difficulty. Toddlers perceived by their mothers as more difficult are also perceived as irritable and prone to protest frequently during various daily activities. The correlation of difficulty with TBAQ Activity Level suggests a quality of high motoric arousal, perhaps related to the acting out, or externalizing, component of difficulty. Turning to the TBAQ correlations with the ICQ subscales, the TBAQ Pleasure × ICQ Un sociability and TBAQ Anger Proneness × ICQ Un adaptability correlations are probably due to straightforward content overlap.

Examination of the ICQ Persistence items indicates that two of the three items that define this scale inquire about continuing to play with objects when told to quit or continuing to go somewhere when told to stop. Thus, the finding that TBAQ Activity Level is related to ICQ Persistence is reasonable in terms of item content and does not indicate a lack of discriminant validity for either scale. In fact, these same two ICQ items help define a 24-month ICQ factor labeled “unstoppable” (Bates, 1992). Similarly, the finding that TBAQ Anger Proneness is related to ICQ Persistence is reasonable because all three ICQ Persistence items concern situations in which the child and parent engage in conflict and the child protests or disobeys. The lack of correlation between TBAQ Interest/Persistence and ICQ Persistent is also completely expectable, given the lack of common content. This detailed content analysis illustrates that one can easily be misled by scale names alone in evaluating patterns of convergence and discrimination.

Relation to parentally reported socioemotional individuality.—Part of the significance of temperament is due to its effects on social interaction (Brody, Stoneman, & Burke, 1987; Munn & Dunn, 1989; Rubin & Asendorpf, 1993). Full investigation of this issue requires observational research, but an initial approach is possible within this article’s defined context of caregiver report by asking how the TBAQ scales relate to parents’ perceptions of their children’s expression of more social emotions. Four questionnaire scales measuring “socialization-relevant traits” were available from Rothbart, Ahad, and Hershey (1994), and I administered these scales to mothers in Sample 8b (see lower portion of Table 4). These scales were Aggression (sample item: “Takes toys away from other children”); Empathy (sample item: “Is upset by stories in which the characters are hurt or die”); Guilt/Shame (sample item: “Doesn’t act upset when s/he has done something wrong”); and Help seeking (sample item: “Usually asks for help on difficult tasks”).

The literature on emotional development allows some predictions. These predicted correlations (underscored in the table) were derived from straightforward content similarity, from relation to a common dimension such as approach/withdrawal or externalizing versus externalizing, or from previous findings. All the predicted correlations were significant with the age 4 TBAQ scales, and most were not statistically significant—but all were in the predicted direction—with the age 18 month TBAQ scales. The correlation of the Aggression scale with TBAQ Activity Level is predictable from the externalizing content in both scales, and TBAQ Anger Proneness and Aggression have common emotional cores. Empathy is perhaps related to a positive versus negative affect (Pleasure vs. Anger Proneness) dimension because both positive affect and empathy are components of prosocial behavior. The link between Social Fearfulness (conceived as inhibition) and Guilt/Shame has been explicated theoretically and demonstrated empirically by Kochanska (1991, 1993). There were several other significant findings that I did not predict. In retrospect, it is quite understandable that toddlers who are not persistent are later more likely to seek help (rather than pursu ing tasks on their own). A reciprocal relation between TBAQ Pleasure and Aggression is similarly understandable. Thus, the temperamental dimensions tapped by the TBAQ are related to several aspects of social behavior. These relations highlight the potential functional significance of temperament and encourage more detailed investigation to eliminate the confounding by the common source of maternal report.

Correspondence between Mothers and Fathers

Three samples have produced data, still limited, on parental correspondence. In Sample 2, when only the Social Fear, Pleasure, and Activity Level scales were available, correlations for parental agreement ranged from .20 to .41. Subsequent studies used Sample 9 (reported by Crowley et al., 1991) and Sample 11b; none of the mother-father correlational differences between Sample 9 and 11b even approached significance, thus justifying computation of the following weighted mean parental agreement
correlations: Activity Level (.52), Pleasure (.29), Social Fearfulness (.42), Anger Proneness (.54), and Interest/Persistence (.30). These three data sets yield several tentative conclusions. First, temperament scales reveal some relationship and/or contextual specificity, but mothers’ and fathers’ judgments are by no means independent. The degree of father-mother agreement falls within the range noted in reviews of the temperament literature (Martin & Halversen, 1991; Slabach et al., 1991). The values tend to be somewhat higher and more consistent across scales (particularly in Sample 9) than many others in this literature. Although this positive outcome might be due to the more explicit situations and responses described in the TBAQ items, I cannot rule out the possibility that the fathers who chose to participate in this study self-selected for more involvement with their toddlers.

Longitudinal Stability
Temperament develops. This consensus (Rothbart, 1989) changed the theoretical implication of analyses of longitudinal stability. Rather than demonstrating stability to document that a trait was indeed temperamental, longitudinal analyses elucidate the patterns of change and continuity as emotional expressivity is transformed and as an array of emotion regulation strategies begin to operate during development. Nevertheless, stability within periods between major transitions in emotional development (which, of course, occur at different times for different features of temperament) is still a strong expectation of temperament.

In Sample 8b, I administered the TBAQ at ages 18 months and 4 years. Table 5 shows the stability of TBAQ scales over this interval. Because age 4 years is at the extreme boundary of applicability of the TBAQ and 2.5 years is an interval encompassing appreciable real developmental change, the values in Table 5 are minimal estimates of the TBAQ’s stability. No very short-term test-retest stability estimates are yet available. The overall picture is one of relative specificity. The underscored stability estimates on the diagonal are moderate, with the exception that toddlers do not maintain their rank order on Social Fearfulness. Two of the off-diagonal correlations suggest, however, that 4-year-olds who are perceived as relatively shy were low in interest and persistence (a finding that fell short of significance) and not anger prone at age 18 months. A significant off-diagonal longitudinal correlation is the association of early anger with later high activity (r = .38), an association observed in other contexts in the TBAQ construction and validation process.

Relationship of IBQ and TBAQ

Overview.—The TBAQ is the obvious follow-up instrument for a sample previously tested with Rothbart’s (1981) IBQ. The corresponding scales are as follows (IBQ scale/TBAQ scale): Distress to Limitations/Anger Proneness; Smiling and Laughter/ Pleasure; Duration of Orienting/Interest/Persistence; Activity Level/Activity Level. The IBQ Distress to Novelty and TBAQ Social Fearfulness scales are not closely matched conceptually because TBAQ Social Fearfulness has substantially more shyness-related content than IBQ Distress to Novelty. There is no TBAQ counterpart to the IBQ Sootability scale. Despite the corresponding scale names, the TBAQ items were not derived from the IBQ items, with a few exceptions. The scale content varies in an age-appropriate fashion, and the TBAQ contains a greater variety of item content.

Table 5

<table>
<thead>
<tr>
<th>18-Month TBAQ</th>
<th>Activity Level</th>
<th>Pleasure</th>
<th>Social Fearfulness</th>
<th>Anger Proneness</th>
<th>Interest/Persistence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity Level</td>
<td>.43*</td>
<td>- .16</td>
<td>-.05</td>
<td>.15</td>
<td>-.14</td>
</tr>
<tr>
<td>Pleasure</td>
<td>.02</td>
<td>.41*</td>
<td>-.20</td>
<td>.04</td>
<td>.16</td>
</tr>
<tr>
<td>Social Fearfulness</td>
<td>.15</td>
<td>.10</td>
<td>.06</td>
<td>-.13</td>
<td>-.16</td>
</tr>
<tr>
<td>Anger Proneness</td>
<td>.38*</td>
<td>-.04</td>
<td>-.31*</td>
<td>.39*</td>
<td>-.15</td>
</tr>
<tr>
<td>Interest/Persistence</td>
<td>-.16</td>
<td>.23</td>
<td>-.25</td>
<td>-.08</td>
<td>.54*</td>
</tr>
</tbody>
</table>

* Sample 8b; N = 37.
* p < .05.
IBQ-TBAQ interrelations.—Some empirical light was shed on this issue by asking mothers with children in the transitional age range—mostly 15, 16, and 17 month-olds—between the IBQ and TBAQ (Sample 10) to complete both questionnaires on one occasion. The resulting significant correlations (all at $p < .05$) for comparable scales follow this list of corresponding scales (TBAQ first, IBQ second): Activity Level, Activity Level, $r = .68$; Pleasure, Smiling and Laughter, $r = .63$; Social Fearfulness, Distress to Novelty, $r = .41$; Anger Proneness, Distress to Limitations, $r = .46$; and Interest/Persistence, Duration of Orienting, $r = .60$. Thus, the observed correlations were highest for Activity Level, Pleasure, and Interest/Persistence, and appreciable for the other two scales. Given the earlier comments about the lack of comparability of the IBQ and TBAQ fear measures, I scored only the five IBQ items that inquire specifically about reactions to strangers. This short miniscale correlated .68 with the TBAQ Social Fearfulness scale, which is notably higher than the .41 correlation observed for the full scales. This confirms that the nonsocial content of the IBQ is responsible for the lower correlation of the full fear scales. The pattern of correlations suggests that the IBQ and TBAQ tap somewhat different aspects of these traits, and this is particularly true for the anger-related scales. The chief difference in anger-related item content is that the TBAQ inquires about social contexts in addition to the feeding, sleeping, and solitary play contexts that are in the IBQ item pool. In addition to this difference in contexts sampled, the TBAQ also inquires about angry instrumental responses in addition to the crying/fussing responses that dominate the IBQ.

These univariate correlational results are obviously crucial in interpreting longitudinal correlations of the IBQ and TBAQ. They set an upper limit to such longitudinal correlations, and, of course, this upper limit is lower than that set by reliability.\(^4\)

Longitudinal analyses.—Ultimately, understanding the stability of questionnaire-based assessment of early temperament will require large-sample longitudinal studies that span the infant, toddler, and preschooler periods. However, the IBQ–TBAQ association is of both practical and theoretical interest. To address this issue, two longitudinal analyses have employed both the IBQ and TBAQ. The first, preliminary study used Rothbart’s IBQ at age 9 months and the early three-scale version of the TBAQ at 22 months (Sample 2). Results indicated moderate stability of mothers’ reports from 9 to 22 months of age for the Social Fearfulness and Pleasure scales, but not the Activity Level scale. A second study (Sample 11a) involved administration of the IBQ at age 12 months and the full TBAQ at 18 months (Table 6).

As Table 6 shows, the stability (across time and instruments) ranges narrowly from .41 to .52. This moderate degree of stability should be understood in the context of the IBQ–TBAQ correlations on a single occasion, which were described above. For example, the observed 12- to 18-month, IBQ-to-TBAQ Anger correlation of .52 (Table 6) might be viewed as only moderate in size. However, it is remarkably high in light of the observed same-age IBQ-TBAQ Anger correlation of only .46. The implication is that what is common to IBQ and TBAQ anger is extremely stable from 12 to 18 months. The same is true in the shyness domain. In fact, all the stability correlations are impressive from this perspective.

Attending to the diagonal correlations in Table 6, there is considerable specificity in the longitudinal relations, a finding that generally supports the case for discriminant validity of both questionnaires in this context. The exceptions to specificity are also informative. That is, six of the 20 off-diagonal stability correlations are significant, and five of these six pertain to positive or negative hedonic tone (pleasure and interest [and soothability in the case of the IBQ] vs. anger and fear); the sixth concerns the previously observed activity-anger association.

Standardization

Reliability and validity data will continue to accumulate for the TBAQ, but the

\(^4\) To derive predicted IBQ scores from TBAQ scores (or vice versa), a regression approach is required. Analyses predicting each TBAQ scale from all the IBQ scales and vice versa were conducted, and interested researchers may request these results from the author. The "corresponding" scale in the other questionnaire tended to carry all the predictive power, and thus the correlational findings provided in the text convey the important substantive findings for the general reader.
TABLE 6
LONGITUDINAL ANALYSES RELATING 12-MONTH IBQ AND 18-MONTH TBAQ SCALES

<table>
<thead>
<tr>
<th>12-MONTH IBQ</th>
<th>Activity Level</th>
<th>Pleasure</th>
<th>Social Fearfulness</th>
<th>Anger Proneness</th>
<th>Interest/Persistence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity Level</td>
<td>.44*</td>
<td>.06</td>
<td>.14</td>
<td>.23</td>
<td>-.18</td>
</tr>
<tr>
<td>Smiling and Laughter</td>
<td>.16</td>
<td>.40*</td>
<td>-.08</td>
<td>-.12</td>
<td>.05</td>
</tr>
<tr>
<td>Distress to Novelty</td>
<td>.21</td>
<td>-.04</td>
<td>.43*</td>
<td>.51*</td>
<td>-.15</td>
</tr>
<tr>
<td>Distress to Limitations</td>
<td>.32*</td>
<td>.06</td>
<td>.56*</td>
<td>.52*</td>
<td>-.26*</td>
</tr>
<tr>
<td>Duration of Orienting</td>
<td>.17</td>
<td>.34*</td>
<td>-.21</td>
<td>-.09</td>
<td>.41*</td>
</tr>
<tr>
<td>Soothability</td>
<td>-.00</td>
<td>.12</td>
<td>-.09</td>
<td>-.24*</td>
<td>.12</td>
</tr>
</tbody>
</table>

Note.—Correlations underscored reflect relations between scales with similar content.

\( p < .05 \)

encouraging results thus far lead us to present standardization data, which can help investigators assign more accurate scores, especially when working with small samples. Table 7 provides means and standard deviations for three groups, which closely agree. Note that means fall toward the middle of the 7-point rating scale, with the exception of Pleasure, which has a higher mean. The distributions tend not to be skewed. It is important to emphasize that these norms are for normal toddlers rated by mothers. Issues related to standardization include gender and age effects.

A goal of item writing for this toddler questionnaire was to avoid gender effects, which are rare in infant temperament research but sometimes present in the preschooler age range. Gender effects on scale means and standard deviations were examined and found to be minimal and inconsistent. I correlated scale scores with gender (coded as a 0, 1 variable) in five samples. The mean correlations were Activity Level, .08 (males higher); Pleasure, -.15 (females higher); Social Fearfulness, .03 (males higher); Anger Proneness, -.02 (females higher); and Interest/Persistence, .11 (males higher). None of these 25 correlations was significant at the \( p = .05 \) level in the individual samples, and none exceeded .24. Considering the average values, the best case for a gender difference is for the TBAQ Pleasure scale, on which girls tended to be rated slightly higher. However, the entire pattern of results suggests that separate norms for females and males are unjustified.

Age effects were also small but more consistent. I examined these effects in the samples that ranged most widely in age (Samples 3, 4, 9a, and 10). There was a tendency for scale scores to increase with age (i.e., for respondents to report that children showed more of the named quality), with the exception of Social Fearfulness, on which children showed less shyness with age in all four samples (mean \( r \) with age = -.14). Of the other four scales, Interest/Persistence

TABLE 7
STANDARDIZATION DATA FOR THE TBAQ

<table>
<thead>
<tr>
<th>Scale</th>
<th>Sample 6 &amp; 7a</th>
<th>Sample 10b</th>
<th>Sample 11c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity Level</td>
<td>Mean 4.26</td>
<td>Mean 4.29</td>
<td>Mean 4.48</td>
</tr>
<tr>
<td></td>
<td>SD .64</td>
<td>SD .68</td>
<td>SD .78</td>
</tr>
<tr>
<td>Pleasure</td>
<td>Mean 5.07</td>
<td>Mean 4.90</td>
<td>Mean 4.87</td>
</tr>
<tr>
<td></td>
<td>SD .62</td>
<td>SD .68</td>
<td>SD .73</td>
</tr>
<tr>
<td>Social Fearfulness</td>
<td>Mean 3.89</td>
<td>Mean 4.22</td>
<td>Mean 4.11</td>
</tr>
<tr>
<td></td>
<td>SD .85</td>
<td>SD 1.04</td>
<td>SD .90</td>
</tr>
<tr>
<td>Anger Proneness</td>
<td>Mean 3.96</td>
<td>Mean 3.82</td>
<td>Mean 4.10</td>
</tr>
<tr>
<td></td>
<td>SD .67</td>
<td>SD .72</td>
<td>SD .79</td>
</tr>
<tr>
<td>Interest/Persistence</td>
<td>Mean 3.82</td>
<td>Mean 3.58</td>
<td>Mean 3.71</td>
</tr>
<tr>
<td></td>
<td>SD .70</td>
<td>SD .85</td>
<td>SD 1.03</td>
</tr>
<tr>
<td>Social Desirability</td>
<td>Mean 4.45</td>
<td>Mean 4.28</td>
<td>Mean 4.31</td>
</tr>
<tr>
<td></td>
<td>SD .43</td>
<td>SD .42</td>
<td>SD .53</td>
</tr>
</tbody>
</table>

\( a^* N = 102 \) (pooled); age 18–24 months.

\( b^* N = 51 \); age 14–21 months.

\( c^* N = 78 \); age 15 months.
Child Development

showed the strongest tendency to increase with age (mean \( r \) with age across the four samples = .23). The mean correlation with age for Pleasure was .17, and the mean correlations with Activity Level and Anger Proneness were quite modest (.10 and .11, respectively). Given these findings, it seems most appropriate to provide standardization data without stratifying by age, with the caveat that age correction could be considered when ages in a sample range widely.

Given that the TBAQ has been translated into Dutch, German, and Japanese, there exists the likelihood that cross-cultural issues can be addressed in the near future.

**General Discussion: Implications for Measuring and Understanding Temperament**

The questionnaire construction and validity results are largely supportive of a traditional view of temperament as recruiting similar responses across situations that have comparable incentive properties and as being relatively stable over limited periods of development. However, these generalizations require qualification and extension because older, stereotyped views that hold temperament to be invariant across relationships and diverse contexts have proven unrealistic.

**Convergence across measures and respondents.**—Some of the data in this report are buttressed by TBAQ data already in the literature. To address the issues of cross-questionnaire convergence and agreement between respondents, Goldsmith, Rieser-Danner, and Briggs (1991) administered the early three-scale version of the TBAQ to mothers and teachers of a sample of toddlers in day-care settings along with the TTS (Ful lard et al., 1984) and the EASI (Buss & Plomin, 1975). The results indicated substantial convergence for conceptually related scales in maternal report data. The TBAQ Social Fearfulness scale correlated strongly with the scale having the closest content on the other two questionnaires, the TTS Approach/Withdrawal scale (\( r = .79 \)) and moderately with EASI Sociability (\( r = -.49 \)). The TBAQ Activity Level scale also correlated substantially with corresponding activity scales on the TTS (\( r = .73 \)) and EASI (\( r = .54 \)). Convergence for activity (but not fear) scales was also apparent in day-care teachers' data. Combined with the results reported in the present article for convergence with Bates's ICQ, it is reason-

able to conclude that the TBAQ shows adequate convergent validity with the three English-language toddler temperament questionnaires that have been most widely used. However, few of these correlations are so high that the scales could be considered interchangeable. The important issue of convergence with laboratory-based measures is under investigation with Goldsmith and Rothbart's (1991) Laboratory Temperament Assessment Battery.

Goldsmith et al. (1991) also reported mother-teacher agreement for the TBAQ Social Fearfulness, Activity Level, and Pleasure scales to be low and nonsignificant; the same null results obtained for most of the TTS and EASI scales as well. How is this result reconciled with the moderate and significant mother-father correlations reported above? There are some uninteresting possibilities of random sampling error, sample selection, or how the different kinds of respondents view their implicit relationship to the researchers. There are also other more implicative possibilities. From the measurement perspective, there is the possibility that mothers and teachers use different reference groups of children in making ratings, whereas mothers and fathers are more likely to know the same other children. However, the TBAQ item format does not call for comparisons, so this explanation would have to involve implicit comparisons. There is also the possibility that the lower teacher-mother agreement is due to teachers having little or no chance to observe the child in certain TBAQ-specified contexts. However, Goldsmith et al. (1991) tested for this explanation by deleting such items and found the explanation wanting. From a theoretical perspective, there exists the possibility that the toddler's actual behavior differs systematically in the home and day-care settings, and this would account for the discrepancy. A more specific version of this contextual specificity explanation is that some behavior reported on temperament questionnaires depends on toddler-caregiver interaction or on the history of that interaction. These interactions are partially conditioned on adults' goals, and because teachers' and parents' interactions with the toddler have different goals, the patterns of their interaction are likely to differ as well. Thus, if respondents typically interact with a child in different situations, or with different goals, then reports of temperamental behavior are likely to reflect those differences, as argued by Goldsmith, Bradshaw, and Rieser-Danner (1986). The
question of contextual specificity in temperamental expression is likely to resist easy resolution. It will require detailed investigation of each domain of temperament.

Continuity and change.—The TBAQ stability results should be understood in the broader context of recent evidence for the continuity of temperament from the late toddler period (age 3) to early adulthood. Threads of continuity from 3 to 18 years are apparent in a representative cohort of over 800 New Zealanders, even when the age 3 assessments were ratings by examiners and the age 18 assessments were by self-report (Caspì & Silva, 1995). A much smaller sample with more intensive, multimodal assessments also demonstrated notable continuity from age 3 to 23 years (Block, 1993). The continuity over these longer periods obviously does not involve precisely the same behavioral patterns, but conceptual links between the actual behaviors assessed at the beginning, intermediate, and ending age periods are apparent. This continuity might reflect an "emotional core" of personality traits that is elaborated both by interpersonal interactions and individual development in the cognitive sphere. Although no long-term evidence concerning the subsequent personalities of toddlers assessed with the TBAQ is in the offing, Goldsmith et al. (1994) reported parent-offspring data for Sample 2 subjects (assessed with the IBQ and later, the TBAQ) whose parents reported about their own personalities using Tellegen's (1982) MPQ. In an independent study also reported in the same paper (Goldsmith et al., 1994), parent personality was assessed using Big Five trait adjective self-ratings, and toddler-aged offspring were again rated on the TBAQ. Positive affect in the parent and toddler ratings were associated (parental extraversion and offspring TBAQ Pleasure), and parental negative emotionality (in the form of low agreeableness) was associated with TBAQ Activity Level and Anger Proneness. Together with the longitudinal research of Caspi and of Block, cited above, these results imply that some dispositions toward adult personality may already be apparent in toddlers.

No consideration of continuity is complete without discussing change. One implication of our results is that stability of temperament tends to be only moderate, and this is consistent with the contemporary view that rigid stability is an inappropriate expectation of temperament. There is perhaps some consensus in the temperament literature that stability is only to be expected within periods bounded by behavioral reorganization. An example of such a reorganization from the well-studied domain of inhibition is when children begin to experience fear in response to the anticipation of social evaluation, whereas earlier fearful feelings were confined to more objectively threatening or novel incentives (Kagan & Snidman, 1991). Contributors to a volume edited by Rubin and Asendorpf (1993) document multiple developmental pathways of fearfulness, inhibition, and shyness in early childhood. Thus, the lack of homotypic stability for Social Fearfulness might be due to the emergence of concern about social evaluation between the ages of assessment in this study. This concern about social evaluation continues and forms a cornerstone of the two-factor view of inhibition (Asendorpf, 1990, 1993). Less is known about the early developmental course of temperamental domains other than fear or inhibition, but similar factors might explain change in pleasure, anger, and interest scores. Of course, such theoretical license opens the possibility that any observed instability can be attributed to ill-understood behavioral reorganizations. Investigators should only view instability of temperament as theoretically acceptable when the occasions of measurement bridge a reorganization that can be specified.

The structure of temperament.—The scale construction process yielded several insights into the structure of temperament. The Anger Proneness scale was salient in several analyses. First, despite efforts to enhance discriminant validity, Anger Proneness was associated with Activity Level, undoubtedly because angry behavior often involves movement and arousal, in contrast to fearfulness. A finding from both our mother-father comparisons as well as a previous study of employing Rothbart's IBQ (Goldsmith & Campos, 1990) was that parental agreement is higher for the anger and activity scales than for others. It is noteworthy that Bates (1992) reported that the IBQ Distress to Limitations (Anger) scale was most closely related to difficulty on the 6-month version of the ICQ. Thus, the same interpretation of difficulty that I proposed for toddlers (i.e., essentially, irritability or anger expression) may also hold for young infants. The salience of difficulty may account for the parental agreement on typical frequencies of anger.

The association of anger and activity is not the only feature of the structure of tem-
perament suggested by these analyses. The expressions of pleasure and interest are also positively correlated, perhaps because expression of pleasure depends on engagement with the environment and/or because Interest/Persistence taps a milder aspect of the same broad hedonically positive dimension as Pleasure. Thus, the TBAQ results are broadly compatible with a “hedonic tone” interpretation of temperament, having dimensions of positive affect and negative affect (Tellegen, 1985). On the other hand, the TBAQ also distinguishes clearly between anger and fear (i.e., they are distinct within the TBAQ scale structure itself, across time, and across different instruments). The results are also compatible with an approach/avoidance interpretation; for example, the best marker of avoidance (Social Fearfulness) is negatively correlated with the best marker of approach (Pleasure) in several analyses. Finally, within the traditional fear domain, I found that items concerning fears in the inanimate world (objects, novel or intense sensory events, and animals) did not cohere with items concerning social fear in the scale construction process. In summary, although the TBAQ is most closely aligned with a discrete affect approach to emotional individuality, other ways of structuring the temperament domain are compatible with the empirical findings.

Some psychometric implications.—Considerable clarity was obtained in these analyses by employing multiple replications rather than depending on statistical significance for inference. In several instances among the 11 samples, isolated, significant associations emerged. Replications tempered the interpretation of these findings. The opposite situation also occurred. Non-significant findings of modest effect size occurred so frequently that confidence in them accrued; an example is the correlation between anger and activity level in several contexts. The need to replicate small effect sizes is particularly crucial in the area of socioemotional development, where multiple influences of modest size are probably the rule rather than the exception.

Another measurement implication concerns the comparable psychometric properties of each TBAQ scale. In contrast to other noniterative methods of scale construction, which tend to yield scales of differing reliability (e.g., factors extracted later vs. earlier in factor analysis, scales with fewer vs. more remaining items after a single round of item analysis), the TBAQ scales are almost equally reliable. Thus, differences in their empirical associations are more likely to reflect substantive processes.

Another measurement implication of the research leading to the TBAQ concerns the relative benefits of different item types. The more general type of item used on some other questionnaires (e.g., “child is active”) implicitly assumes that the respondent recalls and appropriately generalizes over the relevant situations and responses, whereas the more specific type of item, as used in the TBAQ, allows the researcher to extract a score with general meaning. Although the approach with specific wording has the appeal of clear behavioral referents and might appear less susceptible to bias, the relative superiority of the two item types will ultimately be settled empirically.

References
tions of their children from age 6 months to 3 years. Merrill-Palmer Quarterly, 30, 111–130.
Block, J. (1993). Studying personality the long way. In D. Funder, R. D. Furke, C. Tomlin-
son-Keasy, & K. Widaman (Eds.), Studying lives through time: Personality and develop-
ment (pp. 9–41). Washington, DC: APA Press.
dale, NJ: Erlbaum.
velopment. In M. M. Haith & J. J. Campos (Eds.), P. H. Mussen (Series Ed.), Handbook of child psychologi-
Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. Brain and Cogni-
tion, 20, 125–151.
Faddy, W. (1993). Constructing questions for inter-
views and questionnaires. Cambridge: Cambridge University Press.
Fox, N. A. (Ed.). (1994). The development of emotion regulation: Biological and behavioral con-
ity in developing emotion systems. In M. Lewis & J. M. Haviland (Eds.), Handbook of emotion. New York: Guilford.
Goldsmith, H. H., Bradshaw, D. L., & Rieser-
Danner, L. A. (1986). Temperamental dimen-
Goldsmith H. H., Buss, A. H., Plomin, R., Roth-
Goldsmith, H. H., & Campos, J. J. (1986). Funda-
mental issues in the study of early temperament: The Denver Twin Temperament Study. In M. E. Lamb, A. L. Brown, & B. Rog-
goff (Eds.), Advances in developmental psychologi-
Goldsmith, H. H., & Campos, J. J. (1990). The structure of infant temperamental disposi-
ty: A twin study of the five factor model and parental-offspring analyses. In C. Halverson, R. Martin, & C. Kohlstam (Eds.), The develop-
ring structure of temperament and personality from infancy to adulthood (pp. 241–
Goldsmith, H. H., Reilly, J., Lemery, K. S., Long-
ment Battery (version 1.0). Technical Report, Department of Psychology, University of Wisconsin—Madison.
234 Child Development


(Eds.), *Explorations in temperament* (pp. 205–234). New York: Plenum.