Discussion

Behavior genetics studies of infant temperament: Findings vary across parent-report instruments

Julie Hwang, Mary K. Rothbart

Psychology Department, 1227 University of Oregon, Eugene, OR 94703-1227, USA

Received 30 August 2002; accepted 30 August 2002

In a detailed review of issues in the assessment of children’s temperament, Rothbart and Bates (1998) concluded that “temperament measures remain both under construction and under attack—and those who enter this construction zone must wear conceptual hard hats” (p. 127). Donning these hats and hoping they do not provide us with too much protection, we comment on Saudino’s (2003) response to our commentary on Vaughn, Taraldson, Crichton, and Egeland (1981). Although we agree that results indicating “too low” DZ twin correlations in a number of studies raise questions about the possibility of contrast effects in parent judgments, we: (a) provide evidence that these “too low” correlations are found for some instruments but not for others, and (b) suggest that the nature of the judgments required in different parent-report instruments may make them more or less subject to contrast effects.

Saudino’s (2003) discussion paper cites results of “too low” DZ correlations as evidence of parental biases in ratings of infant and child temperament. Data from twin studies have often yielded very low or negative DZ correlations, less than half of MZ twin correlations, and this has been interpreted as suggesting contrast effects (e.g., Neal & Stevenson, 1989). DZ twins share on average 50% of additive genetic variance, and the anomalous pattern of correlations in parent ratings of co-twins has led some to suggest that this is due to parents exaggerating the differences between DZ relative to MZ twin pairs.

It is important to note, however, that the most widely used temperament measures in behavior genetics studies have been the EASI (Emotionality, Activity, Sociability, Impulsivity; Buss and Plomin, 1975) Temperament Survey, the EAS, and the Colorado Childhood Temperament Inventory (CCTI). The EASI and EAS were developed by Buss, Plomin, Rowe, and colleagues, and the CCTI added items from the NYLS approach (Thomas & Chess, 1977).
These instruments measure broadly defined temperament dimensions, with items assessing global judgments of a child’s behavior.

Other parent-report instruments have also been used in behavior genetic research, however, including the IBQ (Infant Behavior Questionnaire; Rothbart, 1981) and TBAQ (Toddler Behavior Questionnaire; Goldsmith, 1986). These questionnaires have not shown the anomalous pattern (Goldsmith, Buss, & Lemery, 1997; Goldsmith, Lemery, Buss, & Campos, 1999). Intraclass correlations for DZ twins on the IBQ ranged from .28 to .61 (mean .44) and MZ twins from .52 to .75 (mean .65), and on the TBAQ, .31–.61 (mean .38) for DZ’s and .54–.82 (mean .70) for MZ’s.

The presence of “too low” correlations in one set of measures, and absence in the other, may be related to differences between the instruments in item content and format. EAS-type items require global judgments and cover a broad range of ages (up to 9 years of age). The IBQ and TBAQ are age-appropriate measures, with items describing specific concrete behaviors, asking about the frequency of behaviors observed within a period of time (1 or 2 weeks). For example, Goldsmith et al. (1999) noted that an EAS Activity Level item such as “Child is very energetic” is rated on a 5-point scale from “not at all” to “a lot.” In contrast, an IBQ Activity Level item such as, “When put into the bath water, how often did the baby splash or kick?” is rated on a 7-point scale ranging from “never” to “always,” over the past week. The use of frequency judgments within specified contexts may elicit more accurate responses and less potential bias. Global items may elicit the kind of comparisons with other children or siblings described by Saudino (2003).

Concrete items also allow a parent to adequately sample behavior within a recent time frame. Truncating this time frame may influence or bias the parent’s sampling of behaviors (e.g., limiting responses to a 48-hr period; Saudino & Eaton, 1991). Hagekull, Bohlin, and Lindhagen (1984) showed that when parents were instructed in detail on aspects of behavior observation (e.g., infant looking behaviors, mood expressions, intensity, duration), and observed over the same period, correlations between parent and observers in home observations were significant, ranging from .60 to .83. They concluded that with appropriate comparisons, parents can indeed be reliable observers of their infants’ behaviors.

In describing the DZ anomaly, Saudino reported, “This puzzling outcome is particularly evident with, but not exclusive to, rating measures that require parents to make global judgments of their child’s behavior” (e.g., Goldsmith, Buss, & Lemery, 1997). Goldsmith et al. did report nine temperament scales with “too low” DZ correlations, but these scales were all from the 15-scale Children’s Behavior Questionnaire (CBQ; Rothbart & Ahadi, 1994), not the IBQ or TBAQ. The CBQ and IBQ/TBAQ are conceptually related, but they differ in rating format. CBQ items require parents to rate their children’s behaviors in specific contexts, but they do not utilize frequency ratings. For example, an Activity Level item “(My child) is full of energy, even in the evening,” rated on a 7-point scale ranging from “extremely untrue” to “extremely true,” involves a global rating. This distinction may have influenced Goldsmith et al.’s findings for the CBQ.

Recent non-twin research using multiple methods has provided additional support for cautious use of parental ratings of temperament. Asendorpf (1990), for example, has made significant contributions to our understanding of behavioral inhibition by using multiple raters, settings, and methods, including parent reports. Rothbart, Derryberry, and Hershey (2000)
found that infant laboratory measures at 3, 6.5, 10, and 13.5 months predicted parent-rated temperament at 7 years. Many temperament scales were significantly correlated, with \( r \)'s reaching .63. While the sample was relatively small, these findings provide support for both infant laboratory measures and parental report, as well as for stability of some aspects of temperament from infancy to childhood.

Finally, we wholeheartedly agree with Saudino that the issue of validity of measurement is critical in temperament research. More research in this area is needed, possibly with twin and adoption samples using the IBQ or TBAQ. It will also be important to test whether global or non-frequency judgment is critical to the “too low” DZ effects.

References


