Journal of Clinical Child & Adolescent Psychology, 40(1), 54–66, 2011

ISSN: 1537-4416 print/1537-4424 online
DOI: 10.1080/15374416.2011.533404

Routledge
Taylor & Francis Group

Probing the Depths of Informant Discrepancies: Contextual Influences on Divergence and Convergence

Anselma G. Hartley
Department of Psychology, Brown University

Audrey L. Zakriski
Department of Psychology, Connecticut College

Jack C. Wright
Department of Psychology, Brown University

This study examined how a contextual approach to child assessment can clarify the meaning of informant discrepancies by focusing on children's social experiences and their *if . . . then* reactions to them. In a sample of 123 children ($M_{age} = 13.30$) referred to a summer program for children with behavior problems, parent–teacher agreement for syndromal measures of aggression and withdrawal was modest. Agreement remained low when informants assessed children's reactions to specific peer and adult events. The similarity of these events increased consistency *within* informants but had no effect on agreement *between* parents and teachers. In contrast, similarity in the pattern of social events children encountered at home and school predicted informant agreement for syndromal aggression and for aggression to aversive events. Our results underscore the robustness of informant discrepancies and illustrate how they can be studied as part of the larger mosaic of person–environment interactions.

Informant discrepancies are a robust but poorly understood phenomenon in childhood assessment research and practice. Numerous studies have investigated the child and informant characteristics that predict discrepancies (see Achenbach, 2006; De Los Reyes & Kazdin, 2005), but our understanding of underlying mechanisms remains incomplete. Many have theorized that such discrepancies could result from differences in children's social environments (e.g., home vs. school) and therefore in the behaviors informants observe (Achenbach, McConaughy, & Howell, 1987). Relatively few studies have directly tested these claims (De Los Reyes, Henry, Tolan, & Wakschlag, 2009) or probed the assumptions on which widely used trait or "syndromal" measures are based (Cervone, Shadel, & Jencius, 2001). The child assessment literature continues to be influenced by a nomothetic trait paradigm in which consistency over ratings remains the theoretical expectation, discrepancies are often considered noise, and thus aggregating information from multiple informants' reports is thought to be the solution to improving the signal (Barkley, 1988; Kraemer et al., 2003; Roberts & Caspi, 2001). Recent research on informant discrepancies (Beck, Hartos, &

This research was supported by award number R15MH076787 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. We are deeply grateful to the parents, teachers, children, staff, and administrators whose cooperation made it possible to collect the data reported here. We would especially like to thank Harry Parad, Director of Wediko Children's Services, and Diana Parad, Director of the Wediko Boston School-Based Program, whose support made this work possible. We acknowledge the invaluable contributions of Stephanie Cardoos, who coordinated the Wediko Transitions Project, the parent project for this study. Thank you also to Sophia Choukas-Bradley and Lindsay Metcalfe for their devoted research assistance.

Correspondence should be addressed to Anselma G. Hartley, Department of Psychology, Brown University, 89 Waterman Street, Providence, RI 02912. E-mail: anselma_hartley@brown.edu

Simons-Morton, 2006; Drabick, Gadow, & Loney, 2008; Guion, Mrug, & Windle, 2009), coupled with related developments in the field of personality (Fournier, Moskowitz, & Zuroff, 2008; Mischel, 2009), encourages us to consider other possibilities: that informant discrepancies reflect meaningful variation in children's behavior across situations and that alternative approaches to assessment are needed to incorporate this variation.

Past studies of informant discrepancies have often relied on standardized syndromal instruments that ask a respondent, typically a parent or teacher, to rate how often a child displays various behaviors over a period of time (e.g., Teacher Report Form [TRF] and Child Behavior Checklist [CBCL]; Achenbach & Rescorla, 2001). The popularity of such instruments stems partly from their efficiency; a short inventory can survey a range of behavior problems. Such assessments can be useful when overall rates of problem behaviors are the main concern (e.g., screening in clinical settings) and when functional origins of the behavior are not a central question. Nevertheless, the emphasis on efficiency involves trade-offs. In essence, these measures adopt an "act frequency" (Buss & Craik, 1983) or aggregationist (Roberts & Caspi, 2001) approach in which behaviors are aggregated to capture the person's "act trend" over a period of observation, usually without specifically examining the situations in which the acts occur. In this view, situational variance in behavior is "filtered out" to measure the individual's true disposition (Barkley, 1988). One form of filtering takes place in the mind of the rater when making a frequency judgment about an act statement (e.g., "hits," "teases"). A second form occurs when various items, including some that provide contextual cues (e.g., "disobedient at school"), are aggregated into broad summary scales or syndrome scores.

Alternative approaches more explicitly incorporate context into the study of personality by examining "*if . . . then*" relationships between social events and people's behavioral responses to them (Mischel & Shoda, 1995; Wright & Mischel, 1987). Rather than emphasizing what a person does on average, contextual approaches focus on the conditional probabilities of how a person reacts under relevant conditions, or p(Behavior | Event). For example, Vansteelandt and Van Mechelen's (1998) study of hostility examined several antecedent events (e.g., *if* frustrated, ignored) and adults' reported hostile reactions (e.g., *then* attack, curse). Personality is thus revealed partly in individuals' behavioral "signatures," or the patterning of behavior around their mean level and across contexts (Fournier et al., 2008; Smith, Shoda, Cumming, & Smoll, 2009; Zakriski, Wright, & Underwood, 2005). Instead of implying that consistency across situations should be generally high, these models suggest that it varies with the similarity of the situations in which people are observed (Mischel & Shoda, 1995). Attention to circumscribed consistencies and to contextualized behavior signatures has enhanced our understanding of individual differences in several domains, including anger, anxiety, and dominance (Endler, Parker, Bagby, & Cox, 1991; Fournier et al., 2008; Van Mechelen & Kiers, 1999).

With these issues in mind, we consider some implications of using syndromal assessments to study informant discrepancies. First, several studies suggest that by focusing on overall frequencies, syndromal assessments conflate dispositional and environmental influences on behavior (Smith et al., 2009). Researchers have noted that children who are nomothetically similar (e.g., high overall levels of anxiety) may be psychologically distinct (Haynes, Mumma, & Pinson, 2009; Scotti, Morris, McNeil, & Hawkins, 1996). For example, one boy may act aggressively when peers tease him but may be teased rarely; another boy may be unlikely to act aggressively when teased but may be teased often. Syndromal measures appear to be sensitive to overall behavior output, as they are designed to be, but insensitive to the interaction of person and context variables that contribute to that output (Wright, Lindgren, & Zakriski, 2001).

A second concern about the syndromal paradigm is that it can obscure individual differences in the contextual patterning of behavior. Research on anger and sadness in adults has found individual differences in people's reactions to the same situations that are not revealed by their overall behavior levels (Van Mechelen & Kiers, 1999; Vansteelandt & Van Mechelen, 2006). Related work found that teacher-reported syndrome scores did not distinguish between boys who were equally high in externalizing behavior but had distinct patterns of responses to nonaversive and aversive events in interactions with peers and adults (Wright & Zakriski, 2001). Other work has revealed stability for both mean-level traits aggregated over situations and for disaggregated patterns of behavior across contexts (Fournier et al., 2008). These findings reinforce calls for methods that are sensitive to people's contextualized behavior patterns (Mischel & Shoda, 1995).

Third, to the extent that syndromal assessments conflate environmental and dispositional influences, interpretations can be especially difficult when two syndromal assessments are compared, as in research on informant discrepancies. A discrepancy could occur if a child has different social experiences at home versus school, differs in how she or he responds to those experiences when they occur, or both. Research on gender differences illustrates these issues (Maccoby, 1998). For example, some work has found that gender differences in overall prosocial behavior stem from how often girls versus boys encounter social events rather than how they respond to those events (Zakriski et al., 2005). In contrast, gender differences in overall aggression stemmed

from differences in both event rates and reaction rates. Such ambiguity in the meaning of overall differences is a challenge for research that relies on syndrome scores.

By shifting the focus from overall behavioral output to *if . . . then* relations between contexts and behaviors, a conditional framework helps explain informant discrepancies. If certain aggressive children are primarily aggressive with peers, parent–teacher discrepancies would be expected in cases where parents have few opportunities to observe their child's peer interactions directly. Researchers have speculated that raters may base their ratings on responses to particular situations at home (e.g., when asked to clean up) or at school (e.g., academic or peer challenges; De Los Reyes et al., 2009; Drabick et al., 2008). They have also speculated that parents may base their ratings on a comparatively limited set of parent–child interactions, whereas teachers may base theirs on a broader set of interactions across situations and interactants, including other children. Laboratory research has found that disruptive behavior with the parent was more closely related to parents' prior ratings of disruptive behavior, whereas disruptive behavior with an examiner was more closely related to teachers' ratings (De Los Reyes et al., 2009).

We suggest that informant discrepancies can be clarified by probing two distinct but interrelated processes highlighted in a conditional approach: the probability that a child encounters a given event, p(Event), and the probability of a behavioral reaction given that an event occurs, or p(Behavior | Event). Our study examined parents' and teachers' ratings for children accepted into a summer program for youth with behavior problems. We used a syndromal instrument that is often employed in cross-informant research (CBCL/TRF), which we expected to show the modest informant agreement ($rs \leq .30$) often found in this area (De Los Reyes & Kazdin, 2005). We also used a measure that assesses the likelihood of encountering several social events (e.g., adult instruction, peer provocation) and the conditional likelihood of children's aggressive and withdrawn reactions to each event. We then explored how two sets of factors, the properties of the conditioning events themselves and the frequency with which they are encountered at home and school, predicted the coherence of and agreement between informants' ratings.

We examined four questions. First, we studied how informants' ratings of behavior are influenced when the task identifies the events to which the child is responding. Past work has shown the expected increases in behavioral consistency as antecedent events become more similar, using both adult self-reports (Fournier et al., 2008; Van Mechelen, 2009) and adult observations of children (Mischel & Shoda, 1995; Shoda, Mischel, & Wright, 1993). The latter work defined what we term "event similarity" based on the interactant (peer vs. adult) and the valence of the interactants' behavior (e.g., peer talk vs. tease; adult praise vs. warn). Although these events seem relevant to parents' and teachers' ratings at home and school, whether the event similarity effect generalizes to informant discrepancy research is unknown. Shoda et al.'s (1993) results were based on short observation periods (hourly activities), whereas cross-informant research relies on retrospective ratings over longer periods (typically months). Moreover, the earlier work obtained observations within the same setting, from adults whose general roles with the children were similar. Based on this, and on evidence that teachers are sensitive to social events even when rating children over longer intervals (Wright et al., 2001), we expected the event similarity effect would be present *within* their perspective and setting (school) where role, relationship, and social interactants remain constant. Thus, teachers' ratings of children's reactions to two specific events should be weakly related when those events differ in both interactant and valence, and most strongly related when they share both. Likewise, within their own perspective and setting, we expected parents' ratings to show a parallel event similarity effect.

Second, we tested how the similarity of eliciting events is related to agreement *between* parents and teachers. On one hand, specification of events could reduce differences between parents and teachers in the situations they spontaneously bring to mind when rating behavior (see De Los Reyes & Kazdin, 2005). For example, even when both raters have knowledge of a child's reactions to peers and adults, parents may be likely to bring to mind certain interactions with adults (e.g., a parent), whereas teachers may bring to mind interactions with peers that are more common in classrooms (Wright & Zakriski, 2001). On the other hand, specifying events may have only a modest effect on agreement when the perspectives and settings differ as much as do parents' and teachers'. Children have different relationships with parents and teachers and may behave differently even in response to the "same" situation (e.g., adult instruction) with one adult versus the other (Drabick et al., 2008; Kraemer et al., 2003; Noordhof, Oldehinkel, Verhulst, & Ormel, 2008). Parents and teachers may also interpret the same responses differently; for example, one may encode failure to comply as opposition and the other may encode it as anxiety about task competency (Drabick et al., 2008; Ferdinand, van der Ende, & Verhulst, 2004). On balance, this evidence led us to predict that agreement between parents and teachers would increase with event similarity but that this effect would be weaker than the anticipated effect within rater perspective.

The event similarity effect just described deals with the properties *of* eliciting events, which is distinct from equally important questions about how often children *encounter* those events in their interactions with peers and adults. Some research has examined how cross-informant

discrepancies are predicted by single variables such as the amount of conflict or parental acceptance in the home environment (Grills & Ollendick, 2003; Kolko & Kazdin, 1993). Other work has shown how teachers' frequency ratings of multiple social events (e.g., peer provocation, teacher praise) can be used to clarify certain syndrome groups, but that work did not examine parent–teacher discrepancies (Wright & Zakriski, 2003). The present research integrates and extends both of these approaches. We provide a description of how children's social environments at home and school can be studied as a multivariate pattern of event frequencies, using adult and peer events that parallel the ones we use to study children's reactions. With these patterns of event frequencies, we show how the similarity of children's home and school environments can be operationalized. We then tested our third hypothesis that home–school similarity would predict informant agreement. For both syndromal ratings and event-specific reaction ratings, we expected parent–teacher agreement to be highest when home and school environments were most similar. We expected that our predictive power would increase as more event rates were used to assess the similarity of children's environments and that it would be especially good when we used events involving interpersonal conflict (e.g., peer provocation, adult discipline) to assess home–school similarity.

Finally, we examined the hypothesis that parent–teacher discrepancies emerge in part from the different behaviors they observe, attend to, and perceive as problematic within their perspective and setting (De Los Reyes, Goodman, Kliewer, & Reid-Quinones, 2008; Youngstrom, Loeber, & Southamer-Loeber, 2000). We expected teachers' syndromal ratings to be predicted by a relatively wide range of their ratings of children's reactions to peer and adult events, reflecting the range of teachers' observations at school. We also expected teachers' syndromal ratings to be predicted especially well by their ratings of children's reactions to peer events because of the importance of those reactions to classroom behavior management. Parents' syndromal ratings should be predicted especially well by their ratings of reactions to adult events, particularly adult instruction or discipline, which are important to behavior management at home.

METHOD

Participants

Parents and teachers provided assessments as part of a larger study of children's response to summer residential treatment for behavioral, academic, and social skills problems (see Zakriski et al., 2005, for additional program description). The sample was 52% White, 28% African American, 13% Hispanic, 5% mixed, 1% Asian American, and 1% Native American. Children ranged in age from 8 to 18 ($M = 13.30$ years, $SD = 2.54$); 81 (66%) were boys and 42 (34%) were girls. Teacher assessments were completed by teachers (63.6%), special educators/counselors/therapists (29%), or teaching assistants (7.4%). These raters spent 3.41 hours/day with the student ($SD = 1.91$) over 20.28 months ($SD = 13.38$) in classes that contained 13.32 students ($SD = 5.86$) and 2.07 adults ($SD = .84$). Participant age was not significantly related to interaction time, time known, or class composition.

Procedure

Data were collected before treatment over 2 consecutive years. To ensure that research participation did not affect access to service, informed consent documents were sealed until admissions decisions were made. With most admissions in May and June, it was difficult to reach teachers before the end of school, but we attempted for all 260 children admitted before the last week of June. We obtained complete parent data on 85% of these cases ($N = 222$) and complete teacher data on 60% ($N = 157$). The intersection of these sets ($N = 123$) became the cross-informant sample. Comparisons of these children with those who had only complete teacher data ($n = 34$) or only complete parent data ($n = 99$) yielded no significant differences for age, sex, or TRF/CBCL behavior problems, respectively. Parents and teachers were paid $20 for participation; teachers assessing another child in the 2nd year were paid $30.

Measures

The 118-item CBCL and TRF (Achenbach & Rescorla, 2001) assesses overall behavior problems using eight narrow syndromes and two broad ones (externalizing, internalizing). Ratings are made on a 0-to-2 scale, ranging between 0 (*not true*), 1 (*somewhat or sometimes true*), and 2 (*very true*). We used the Aggression and Withdrawal scales because our contextual assessment focused on these behaviors. Because CBCL/TRF aggression scales differ in item content, we computed alpha coefficients separately. The parent and teacher alphas (see Table 1) did not differ significantly (Feldt & Kim, 2006) and were comparable to those reported by Achenbach and Rescorla (2001). Using their age and gender norms, raw scores were converted to clinical T scores. Means and standard deviations for parents and teachers for CBCL/TRF aggression scales were 70.54 ($SD = 9.85$) and 70.49 (10.85), respectively, and for withdrawal they were 64.49 (10.52) and 62.04 (7.97). Withdrawal means differed, $t(122) = 2.64$, $p < .01$, but not aggression, externalizing or internalizing. Percentages of cases meeting the clinical cutoff (T score ≥ 70) were 52.8% versus

TABLE 1
Internal Consistency and Cross-Informant Agreement for Syndromal Ratings (CBCL/TRF) and for Event-Specific Reactions

		Event-Specific Reactions			
Behavior	*CBCL/TRF*	*Peer+*	*Adult+*	*Peer–*	*Adult–*
		Internal Consistency[a]			
Aggression	.89/.92	.90	.88	.85	.88
Withdrawal	.78/.79	.88	.88	.90	.90
		Cross-Informant Agreement[b]			
Aggression	.23**	.18*	.13	.22*	.18*
Withdrawal	.30**	.10	.14	.20*	.27**

Note: CBCL = Child Behavior Checklist; TRF = Teacher Report Form; Peer/Adult+ = nonaversive peer/adult events; Peer/Adult– = aversive peer/adult events.

[a]Alpha coefficients based on items within each scale.

[b]Correlations (Pearson's *r*) between parent and teacher ratings.

*$p < .05$. **$p < .01$.

41.5% for parent and teacher aggression, respectively, and 26.8% versus 11.4% for withdrawal.

The Behavior-Environment Transactional Assessment (BETA; Hartley, Zakriski, Wright, & Parad, 2009; see Wright & Zakriski, 2001) is a 134-item instrument based on observations of children in treatment (Zakriski et al., 2005). All items are rated on a scale of 0 (*never*) to 5 (*almost always*). We used 48 "reaction" items that assess the likelihood of aggressive or withdrawn behavior if some event occurs. Informants read the overall instruction, "Please rate how this child reacts to the event described," followed by one of eight event prompts (e.g., "If a peer teases or bosses this child . . . "). For each event prompt, they then rate how often the child shows a specific reaction (e.g., "he/she hits, pushes, or physically attacks"). The eight event prompts are "If a peer talks to this child in a friendly or supportive way," "If a peer asks or tells the child to do something," "If a peer argues or quarrels with this child," "If a peer teases or bosses this child," "If an adult talks in a friendly or supportive way to this child," "If an adult gives instructions or directions to this child," "If an adult warns or reprimands this child," and "If an adult disciplines or punishes this child." Aggressive reactions are "argues or quarrels," "whines or complains," "teases or bosses," and "hits, pushes, or physically attacks." Withdrawn reactions are "withdraws or isolates self" and "looks sad or cries." For all analyses, the four aggressive reactions were averaged to form an aggressive reaction composite for each event, as were the two withdrawn reaction items. We also used eight additional "event" frequency items from the BETA. Informants are instructed, "In general, how often do peers or adults do the following things to this child." They then rate each of the eight events just listed (e.g., "Peers tease or boss this child."). The remaining BETA items were not used, either because they assess prosocial reactions not related to the CBCL/TRF or because they assess reciprocal responses to child behavior (e.g., "If this child withdraws . . . " how often do "peers talk in a friendly way to him"?).

"Event similarity". We performed preliminary analyses needed to test how the similarity of specific events was associated with parent–teacher agreement for children's aggressive and withdrawn reactions. Recall that "event" refers to a specific situation on the BETA that might elicit a behavior (e.g., "if a peer teases"). To replicate Mischel and Shoda (1995), we used ratings of reactions to each of the eight events (e.g., withdrawal in response to adult instruction; aggression in response to peer argue; and thus eight reaction scales per behavior). We tested for differences between parents' and teachers' alphas. We set the Type I error rate to .01 because there were 16 comparisons; none was significant. For parents and teachers combined, the alphas for aggressive reactions ranged from .69 (to peer argue) to .85 (to peer talk), with a median of .78. For withdrawn reactions, the alphas ranged from .74 (to adult discipline) to .83 (to peer tease), with a median of .79.

As in Mischel and Shoda (1995), each event has two features: person (adult or peer interactant) and valence (nonaversive or aversive). A similarity index for within-rater agreement of 0, 1, or 2 captures this. For example, peer tease and adult talk share 0 features, peer tease and peer talk share 1 (person), peer talk and adult talk share 1 (valence), and peer tease and peer argue share 2 (valence and person). For parent–teacher agreement, an event can also be "identical" (e.g., "peer tease" for both raters), indexed as a 3. Aggressive (or withdrawn) reactions to events were intercorrelated using Pearson's *r*. For "within-rater" analyses of parents (or teachers), this yielded the lower half of the 8×8 correlation matrix, or 28 *r*s. Each entry in the upper half of the matrix is identical to the corresponding entry in the lower half. Each entry on the diagonal is an autocorrrelation and hence $r = 1.0$. For "between-rater" analyses, this yielded the full 8×8 matrix, or 64 *r*s. Each *r* conveys unique information and each entry on the diagonal is the correlation between parents' and teachers' ratings of children's reactions to "identical" events.

Remaining analyses aggregated aggressive and withdrawn reactions over related events to reduce the number of variables and to simplify the presentation. This resulted in four aggressive reaction scales, each comprising eight BETA reaction items: aggression to peer aversive events (aggression to peer argue combined with aggression to peer tease), to peer nonaversives (peer talk, peer ask), to adult aversives (adult warn, adult discipline), and to adult nonaversives (adult talk, adult instruct). There were four parallel withdrawn reaction scales (each comprising four original BETA reaction items). Four event rate scales were also used: peer/adult aversive event rates and peer/adult nonaversive event rates (each

comprising two BETA event frequency items). As before, we computed alphas for parents and teachers separately and tested for differences. Only two were significant: parent versus teacher ratings of aggression to peer aversives (.80 vs. .89, respectively; $p < .005$), and rates of adult aversive events (.83 vs. .93, $p < .001$).[1] Therefore, Table 1 gives alphas for reaction scales for parents and teachers combined. Alphas for the event rate scales were adult aversives (.88), peer aversives (.70), adult nonaversives (.67), and peer nonaversives (.52). Caution will be needed regarding analyses of nonaversive peer events.

Before standardizing within age and gender (as on the CBCL/TRF), we performed analyses of variance (ANOVAs) on raw BETA ratings to examine age and gender differences. We also examined whether reaction ratings varied over events, as this would clarify whether raters attended to event cues. The mixed-model ANOVAs used age (<12 vs. ≥12 years) and gender as grouping variables, and rater and event type as repeated measures. To avoid overfitting, we restricted the model to main effects and two-way interactions. Greenhouse-Geiser adjustments were used. As expected, the mean level of children's aggressive reactions varied over events, $F(3, 360) = 163.49$, $p < .001$, $\eta_p^2 = .58$, ranging from 1.31 ($SD = .79$; aggression to adult nonaversives) to 2.42 ($SD = 1.00$; to peer aversives). Thus, informants clearly attended to and were influenced by event cues on the BETA. Young children reacted more aggressively than old, $F(1, 120) = 9.09$, $p < .01$, $\eta_p^2 = .07$ ($My = 1.99$, $Mo = 1.61$), especially to peer aversives, $F(3, 360) = 6.60$, $p < .01$, $\eta_p^2 = .05$. Withdrawn reactions also varied in their mean levels over events, $F(3, 360) = 89.84$, $p < .001$, $\eta_p^2 = .43$, ranging from .96 ($SD = .85$; withdrawal to peer nonaversives) to 1.85 ($SD = 1.19$; to adult aversives). The frequency of these eliciting events also varied, $F(3, 366) = 192.78$, $p < .001$, $\eta_p^2 = .61$, ranging from 1.97 (peer aversives) to 4.13 (adult nonaversives). Parents reported more adult aversives (3.01) than teachers did (2.69), $F(3, 366) = 5.35$, $p < .01$, $\eta_p^2 = .04$.[2] To parallel the CBCL/TRF, henceforth we use age- and sex-standardized BETA ratings (z scores).

"Environment similarity." To analyze the similarity of children's social experiences at home and school we used aggregated ratings of how often children encountered peer nonaversive, peer aversive, adult nonaversive, and adult aversive events. Thus, each child had a vector of four event rates as reported by parents and a vector of four event rates as reported by teachers (each standardized within rater). A child's similarity score was the sum of the squared deviations between the vectors, or "Euclidean" distance (Borg & Groenen, 1997). To examine how this environment similarity measure was associated with cross-informant agreement, we then split children at the median into two groups (low vs. high similarity). For each group we calculated cross-informant rs (e.g., for CBCL/TRF aggression or for BETA aggression to adult aversives).

Note that this procedure computes environment similarity using all four event scales. To examine whether the relative breadth of the similarity measure was associated with rater agreement, we also computed environment similarity using less than the full vectors. For example, one possible subset of two events involves parents' (and teachers') ratings of peer aversives and peer nonaversives; another subset involves their ratings of peer aversives and adult aversives. In all, there are six possible subsets of two events per subset. In this way, we identified all possible subsets of one event (there are four subsets), of two events (six subsets), and three events (four subsets). For each subset, similarity was recalculated using the method just described, similarity groups formed using a median split, cross-informant rs computed, and finally averaged within each level of breadth. A second set of analyses calculated similarity for individual event scales (e.g., frequency of peer aversives). Informant agreement was then calculated for children who were more or less similar in how often they experienced that type of interaction. Over the multiple samples generated, the median total sample sizes for low/high-similarity groups were 62/61, respectively.

RESULTS

Syndromal Agreement, Reaction Agreement, and Similarity of Conditioning Events

For comparison with other cross-informant research, we assessed informant agreement using the CBCL/TRF. As expected, agreement (r) between parents and teachers was modest (see Table 1). For comparison, Achenbach and Rescorla (2001) reported rs of .33 and .24 for withdrawal and aggression, respectively. We next examined

[1] We also tested for possible differences between CBCL/TRF scales and the corresponding parent/teacher BETA reaction scales. For aggression, the only difference was that CBCL aggression had a higher alpha than parent BETA aggression to peer aversives (.89 vs. .80, respectively, $p < .002$). For withdrawal, all coefficients for BETA reaction scales were higher than their CBCL/TRF counterparts (the smallest difference .79 vs. .88, $ps < .01$).

[2] Parallel analyses for reactions and events using the eight individual events yielded main effects for events that resembled those reported using four event categories. Briefly, η_p^2 s were .56 for event frequencies, .57 for aggressive reactions, and .43 for withdrawn reactions. We also checked whether the means differed for the pairs of events that were most similar. Every pair differed by Tukey's Honestly Significant Difference (HSD) test, for event frequency ratings (peer argue vs. peer tease), for ratings of aggressive/withdrawn reactions to events (to adult talk vs. to adult instruct), or both (adult warn vs. adult discipline; peer talk vs. peer ask).

the hypothesis that informant agreement would improve when parents and teachers rated children's reactions to specific conditioning events. To facilitate comparisons with results for (aggregated) syndromal ratings just reported, we used aggregated reactions (e.g., aggression to peer aversives). As shown in Table 1, agreement remained modest. All *r*s were positive, but none was significantly different from the *r* for CBCL/TRF agreement.

The preceding results show that informant agreement was modest for reactions to identical events, but they do not provide a complete test of the event similarity hypotheses. First, as we have noted, the similarity of events could play a large role *within* rater or setting (e.g., parent/home) but less of a role when two different settings are involved (home vs. school). Second, although cross-informant agreement was modest for identical events (see Table 1), it could be even lower (or negative) for dissimilar ones. Each of these results would clarify whether and how the event similarity effect applies to research on informant discrepancies.

Figure 1 provides the relevant results. The top panel shows the mean *r*s for aggressive reactions for each similarity level (i.e., number and type of features shared). As hypothesized, agreement within rater increased with event similarity. To summarize this effect, we predicted the pairwise *r*s using 0, 1, or 2 to index event similarity, as previously noted. For parents, *r* increased with similarity, $F(1, 26) = 27.19$, $p < .001$, $R^2 = .51$ (see Figure 1, top). The same was true for teachers, $F(1, 26) = 13.54$, $p < .001$, $R^2 = .34$. Parallel results were found for withdrawal (Figure 1, bottom): *r* increased with similarity within parents, $F(1, 26) = 29.02$, $p < .001$, $R^2 = .53$, and teachers, $F(1, 26) = 12.36$, $p < .01$, $R^2 = .32$.

Figure 1 also examines whether event similarity was related to agreement *between* parents and teachers. We predicted the pairwise *r*s using the between-rater 0, 1, 2, or 3 similarity index (see Method section). As previously noted, 3 reflects pairs for which parents and teachers rated reactions to an "identical" event (e.g., "peer tease"). Event similarity had no effect on cross-informant agreement; all mean *r*s were below .20.[3]

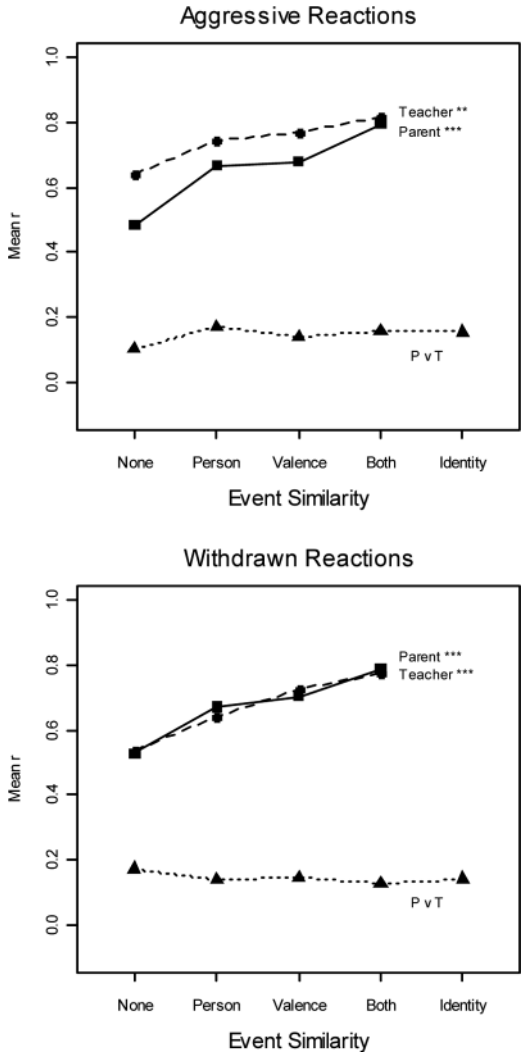

FIGURE 1 Mean correlations among ratings of event-specific aggressive and withdrawn reactions, as a function of event similarity and source of rating. *Note:* Teacher = agreement within teachers' ratings; Parent = agreement within parents' ratings; P v T = cross-informant agreement between parents and teachers; Person = events sharing person type (peer vs. adult); Valence = events sharing valence (aversive vs. nonaversive); None = events sharing neither; Both = events events sharing both; Identity = ratings for identical events by different informants. Asterisks indicate significance of regressions predicting *r* from event similarity: $^{**}p < .01$. $^{***}p < .001$.

Similarity of Experienced Home and School Environments

Clearly, asking parents and teachers to rate children's reactions to events does not mean children encounter those events equally often at home and school. Indeed, cross-informant agreement for ratings of how often children encountered events resembled what is found for syndromal measures: peer nonaversives ($r = .16$, $p < .10$), adult nonaversives (.14, $p < .10$), adult aversives (.24, $p < .01$), and peer aversives (.38, $p < .001$). We have

[3]We checked the possibility that cross-informant agreement for event-specific reactions might be low because some events did not occur often enough for raters to provide a meaningful response. For example, if a rater states that a child is rarely teased, ratings of the child's reaction to teasing may not be informative. To examine this, we repeated the analyses just described, each time removing reaction ratings for which a rater gave the corresponding event a frequency of less than 1, 2, or 3. Cross-informant results were similar to those already reported, with one exception: For aggressive reactions, we found small but reliable increases with similarity, with R^2 of .07, .09, and .14, for the three frequency thresholds, respectively, Fs(1, 54) > 4.89, ps < .04. Cross-informant agreement remained low, with a maximum *r* of .21, at "identity" for the highest frequency threshold.

noted how the similarity of children's experiences at home and school can be defined using rates of events (see Method). We now test the hypothesis that informant agreement will be higher for children whose home and school environments are similar, especially as more events are used to assess similarity.

The results of environment similarity analyses for CBCL/TRF aggression are shown in Figure 2 (top). The cross-informant *r* presented earlier without considering environment similarity is included (see "All Cases"). The abscissa indicates the number of events used to compute environment similarity, or "breadth" (see Method section). For high-similarity groups, cross-informant agreement increased with breadth and was significant at each level. For low-similarity groups, agreement decreased with breadth and was never significant. The differences between high and low similarity groups were significant when 3 or 4 events were used ($zs > 2.16$, $ps < .05$).[4]

The same approach was used for CBCL/TRF withdrawal. There were no significant pairwise differences between the low- and high-similarity groups, regardless of number of events used to compute environment similarity. Although this dictates caution, we note that the only cross-informant correlations that differed from zero were for children with *dissimilar* environments; this occurred at all levels of breadth ($rs = .28$–$.39$, $ps < .05$).[5]

The previous analyses do not examine whether similarity in the rate of encountering aversive events is especially useful in predicting agreement. As shown in Figure 2 (bottom), agreement for CBCL/TRF aggression was higher for the high- versus low-similarity group when environment similarity was based on either peer or adult aversives ($z > 2.02$, $ps < .05$). Thus, even when narrowly defined using rates of aversive events, environment similarity predicted agreement. Recall that these events were the most reliably assessed by parents and teachers.

The same analyses were conducted for withdrawal. Although some individual cross-informant correlations were significant, parent–teacher agreement for withdrawal did not differ significantly for any of the low- versus high-similarity comparisons. Thus, we found little evidence that informant agreement for withdrawal was associated with specific environment similarities.

We performed parallel analyses using ratings of children's reactions to events. As Figure 3 shows, the results for aggression to aversive events resembled those for CBCL/TRF aggression. For multiple-event analyses (top row), environment similarity effects were strongest

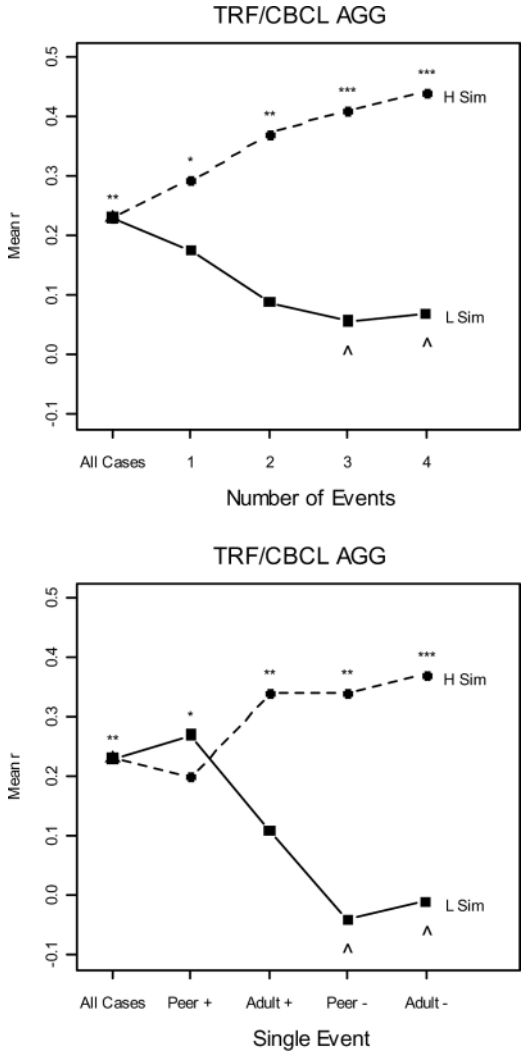

FIGURE 2 Cross-informant agreement (mean *r*) between parents' (CBCL) and teachers' (TRF) aggression ratings (AGG), as a function of number and type of events used to assess similarity of children's home and school environments. *Note:* All Cases = results for all children regardless of environment similarity ($N = 123$); L/H Sim = children with low (L) / high (H) similarity social environments (defined based on median split). Top panel shows agreement as a function of the number of events used to assess environment similarity. Bottom panel shows agreement based on specific events used to assess similarity: Peer/Adult+ = nonaversive peer/adult events; Peer/Adult– = aversive peer/adult events. Tests of *r*/Pairwise tests: $^{*}/\wedge p < .05$. $^{**}p < .01$. $^{***}p < .001$.

[4]One might argue that environment similarity is an indirect measure of children's behavior (e.g., children with similar environments might be more aggressive than those with dissimilar ones). To examine this, we correlated the similarity measure (based on four events) with children's aggregated aggression (i.e., average of TRF and CBCL). Similarity showed little relationship with overall aggression or withdrawal ($rs = .14$, .17, respectively, *ns*). Our results indicate that children with similar environments were more *consistent* in their aggression across settings, but they could be either low or high in their overall aggression (or withdrawal).

[5]To examine heterogeneity within groups, exploratory cluster analyses were conducted. Some children were in the "low-similarity" group because they often had aversive encounters with peers at school but rarely at home. For others, home and school differed primarily in the rates of nonaversive adult events. Children in the "high-similarity" group also showed a variety of event profiles (e.g., high peer aversives at home and school; high adult aversives in both). It will be important in future research with larger samples to examine these functional subgroups.

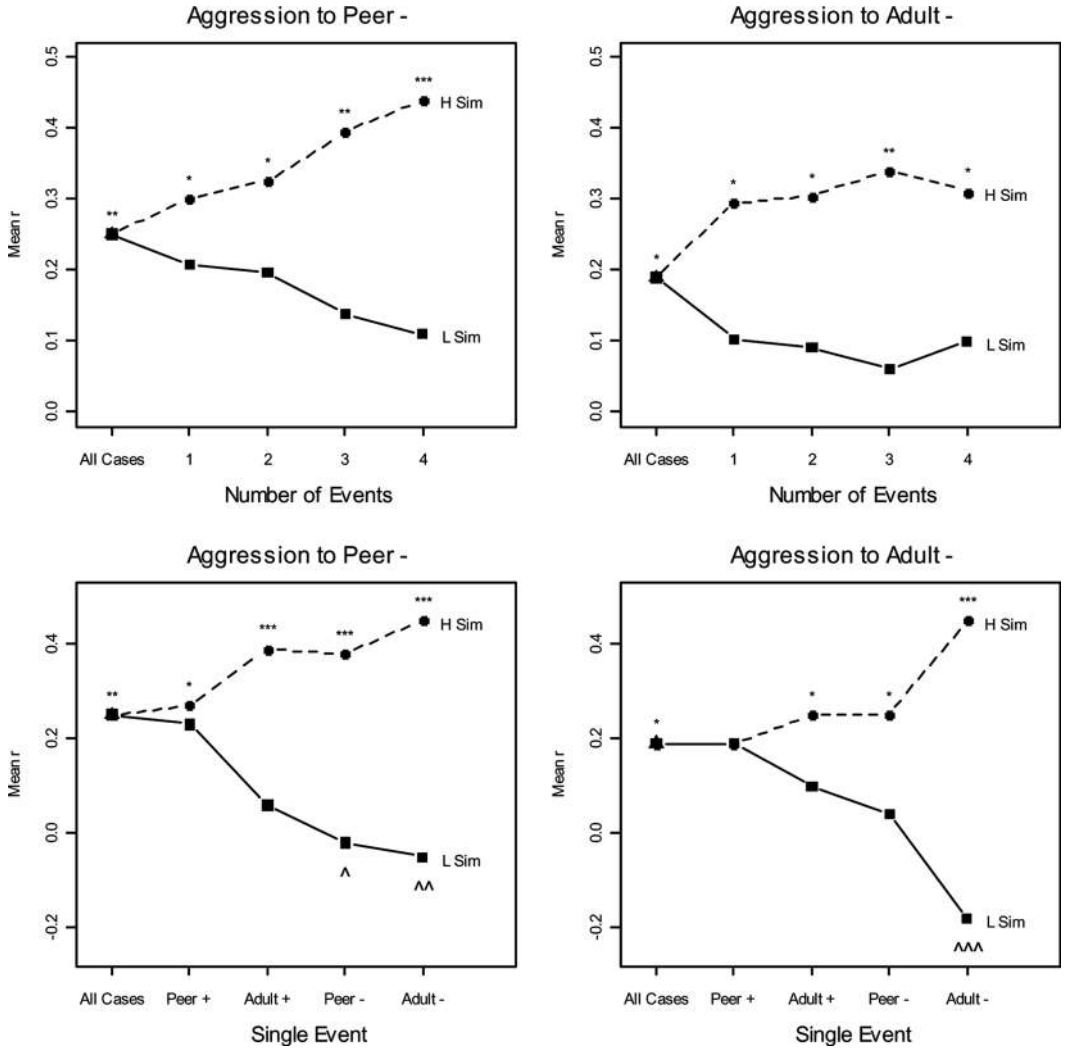

FIGURE 3 Cross-informant agreement (mean *r*) between parents' and teachers' BETA reaction ratings, as a function of number and type of events used to assess similarity of children's home and school environments. *Note:* All Cases = results for all children ($N = 123$); L/H Sim = children with low (L) / high (H) similarity environments (defined based on median split). Top row shows agreement as a function of the number of events used to assess similarity. Bottom row shows agreement based on specific events used to assess similarity: Peer/Adult+ = nonaversive peer/adult events; Peer/Adult − = aversive peer/adult events. Tests of *r*/Pairwise tests: $^{*}/\wedge p < .05$. $^{**}/\wedge\wedge p < .01$. $^{***}/\wedge\wedge\wedge p < .001$.

for reactions to peer aversives. For single-event analyses (bottom), environment similarity effects were clear for aggression to both types of aversives and strongest when similarity was based on adult aversives. Results for withdrawn reactions resembled those for CBCL/TRF withdrawal: Few informant correlations were significant (some for children with similar and some for dissimilar environments), and agreement did not differ between low- versus high-similarity groups.

Predicting Syndromal Ratings from Event-Specific Reactions

Finally, we tested hypotheses noted earlier about the contribution of event-specific reactions to the prediction of parents' and teachers' syndromal ratings. For example, for CBCL aggression, the predictors for the multiple regression were parents' ratings of children's aggressive reactions to each of four events (Table 2). Aggression to adult aversives was predictive for both CBCL and TRF aggression, and aggression to peer positives was also predictive for teachers' ratings of TRF aggression. The vector of four coefficients for parents (row 1) correlated .53 with the vector for teachers (row 2); none of the pairwise comparisons of slopes (parent vs. teacher) was significant. Withdrawal to peer nonaversives was predictive for both CBCL and TRF withdrawal, and withdrawal to adult aversives was also predictive for parents' ratings of CBCL withdrawal; the vectors of coefficients (rows 3 vs. 4) correlated −.18. There was one significant pairwise comparison for withdrawn reactions to adult aversives (warning and

TABLE 2
Regressions Predicting Parents' and Teachers' Syndromal Ratings from Event-Specific Reactions

	Independent Variable: Event-Specific Reaction					
Dependent Variable	*P+*	*P–*	*A+*	*A–*	R^2	*F(4, 118)*
CBCL AGG	.18	.11	.16	.33**	.44	23.37***
TRF AGG	.29*	.18	.07	.28*	.54	35.21***
CBCL WDR	.25*	.11	−.01	.38**	.40	19.46***
TRF WDR	.39**	−.02	.25	.01	.37	17.56***

Note: Predictors for each dependent variable (AGG/WDR) were behaviorally matched reactions. Entries for predictors are standardized regression coefficients. P+ = reactions to peer nonaversives; P– = to peer aversives; A+ = to adult nonaversives; A– = to adult aversives; CBCL = Child Behavior Checklist; TRF = Teacher Report Form; AGG = aggression; WDR = withdrawal.

*$p < .05$. **$p < .01$. ***$p < .001$.

discipline), $t(236) = 2.58$, $p < .02$. The vector of coefficients for parent aggression (row 1) correlated .80 with their vector for withdrawal (row 3); for teachers, the parallel *r* (rows 2 vs. 4) was .01.

DISCUSSION

The current study underscores the robustness of informant discrepancies and illustrates how our understanding of them can be deepened by an analysis of the social events children experience and how they react to them. As expected, parent–teacher agreement for aggression and withdrawal on a widely used syndromal measure was modest ($rs \leq .30$). We predicted that parents' and teachers' ratings of reactions to specific events would show greater agreement, but they did not. The similarity of eliciting events had the expected effect on consistency of reaction ratings *within* a given rater perspective but not on agreement *between* parents and teachers. Agreement was low not only when parents and teachers rated the same behavior in response to different events (e.g., aggression to adult instruction vs. to peer teasing) but equally low when events were nominally identical. In contrast, the similarity of children's social environments, defined in terms of how often children encountered events at home versus school, was linked to informant agreement for ratings of aggression. Predictions of parents' versus teachers' syndromal ratings by event-specific reactions were more similar than expected, but subtle differences emerged.

Our findings demonstrate the reliability and internal organization of raters' contextualized assessments, but they also demonstrate how difficult it is to bridge the gap between informants in different settings. Past research led us to expect that specifying the eliciting event might clarify the assessment task, capture meaningful variability in children's behavior, and at least modestly improve informant agreement. Nevertheless, informant agreement remained low. This finding would be unremarkable if raters ignored the event cues they were given, but they did not. Mean reaction ratings varied considerably over the four event categories (with η_p^2s of .43–.58), showing that raters attended to the conditions that elicit behaviors. Even within the pairs of individual events that were most similar (e.g., adult warn vs. adult discipline), raters distinguished between events, either in terms of their frequency and/or in terms of children's reactions to them. Furthermore, analyses revealed the expected event similarity effect *within* rater. The more features (interactant, valence) two events shared, the greater the consistency of the rated reactions to those events. These results extend past research by showing how the event similarity effect found using direct observations of behavior (Mischel & Shoda, 1995) applies to retrospective ratings within a given setting yet also show that the effect does not apply when parents and teacher rate children's behavior at home versus school.

It is notable that cross-informant agreement remained modest for ratings of reactions to aversive events. Like others (Shoda et al., 1993; Zakriski et al., 2005), we found that problem behaviors were more common and somewhat more variable in response to aversive events (e.g., peer tease, adult discipline), mitigating against floor and restricted range effects. Others have argued that stressful situations engage preferred and stable coping strategies (Hood, Power, & Hill, 2009; Parker & Wood, 2008). Moreover, an aversive "event" (e.g., adult warn) is sometimes a response to a problem behavior previously displayed by the child (Burke, Pardini, & Loeber, 2008). If aggressive children show consistent "bursts" when disciplined (Granic & Patterson, 2006), one would expect better agreement for ratings constrained to this context. These arguments notwithstanding, agreement remained low for reactions to aversive events. The issue is not simply that cross-informant agreement was low in some absolute sense, which researchers have known for some time. Rather, the issue is that parent–teacher agreement was unaffected by factors that one would expect to affect it, and that did affect it within perspective.

Although assessing behavior in response to eliciting events could not bridge the cross-informant gap by itself, attention to how often children *encountered* those events in their interactions was useful. Cross-informant agreement for CBCL/TRF aggression was higher for children whose home and school environments were more similar, and the magnitude of this effect increased as more aspects of children's social experiences were used. Similarity even for single events—namely, conflict with peers or adults—predicted informant agreement. Similarity based on adult instruction and conversation was less useful, and similarity based on peer requests

and conversation was even less so. Reliable assessment of nonaversive experiences, especially for peers, deserves attention in future research. Environment similarity effects were also observed for the already-contextualized reaction ratings. When home and school were more similar, we found better parent–teacher agreement for aggression to adult and peer conflict. Similarity in how often children were disciplined was especially useful in predicting agreement for aggression to that event.

These results are consistent with the view that nominally similar events children experience in everyday social interaction (e.g., ''peer tease'' or ''adult discipline'') can be interpreted differently by informants at home and school (Drabick et al., 2008; Ferdinand et al., 2004). The particular peers and adults differ, and the nuances of their teasing, disciplining, or other actions differ as well. Such variation in the meaning of events is, arguably, part and parcel of the larger phenomena of cross-situational variability and informant discrepancies. Nevertheless, it remains possible that further specification of events would improve informant agreement beyond what we found. It is also possible that parents and teachers believe that children's behavior is cross-situationally variable even when they are given precisely the ''same'' instruction or discipline by different adults. Future laboratory research examining informants' perceptions of children's responses to specific events and interactants would be especially helpful (see De Los Reyes et al., 2009). It is important to note that the goal should be not to reduce informant discrepancies but to measure them well and to understand when and why they do or do not occur.

We found little evidence that cross-informant agreement for withdrawal increased with the similarity of home and school environments. This could be because our sample had fewer clinically withdrawn children, because many of the children with elevated withdrawal were comorbidly elevated for aggression (76% by parent report; 64% by teacher report), or because parents reported higher levels of withdrawal than teachers. Withdrawal was as reliably assessed as aggression for both the CBCL/TRF and the BETA, but for each measure fewer withdrawn behaviors were assessed than for aggression. It is also possible that the social events that predict agreement for withdrawal are more subtle and difficult to assess than are the vivid aversive events we found to be useful in predicting agreement for aggression. This converges with the finding that internal consistency and informant agreement were higher for aversive experiences than for the nonaversives ones (e.g., adult instruction, peer conversation).

Although our home–school similarity measure used a familiar metric (Borg & Groenen, 1997), it should be interpreted with care, especially in light of our ''dissimilarity'' findings for withdrawal. A variety of patterns could (and did) produce comparable (dis)similarity scores. For example, some children who were seen as withdrawn by both their parents and their teachers experienced a specific type of environment *dissimilarity* (e.g., they encountered more peer aversives at school than at home). A detailed analysis of the distinct event patterns experienced by children with low- or high-similarity environments was beyond what our sample size would allow, but this may be a fruitful issue for future research.

Our regression analyses are relevant to discussions about the nature of parents' and teachers' syndrome ratings and the sources of discrepancies between them (De Los Reyes et al., 2008, Drabick et al., 2008; Youngstrom et al., 2000). Parents' and teachers' syndrome scores could be predicted from children's event-specific reactions ($R^2 = 37$–54%), and, as expected, parents' syndromal ratings of withdrawal were predicted better by withdrawn reactions to adult discipline than were teachers'. This effect was narrower than predicted and was not found for aggression. We did not find clear evidence that teachers' ratings were better predicted than parents' by reactions to peers. The overall similarity of parent and teacher regressions dictates caution, but the pattern of findings raised questions that may be worth exploring in future research. One relates to the possible importance of reactions to adult discipline in predicting parents' ratings of problem behavior. Another relates to the role of children's withdrawn reactions to positive peer interactions in predicting teachers' ratings of problem behavior. Interviews and vignette methodologies could be used to further explore how raters prioritize or bring to mind certain interpersonal situations when rating behavior, and whether they sample from the same or different sets of situations depending on the type of behavior they are asked to assess.

Implications for Research, Policy, and Practice

Taken as a whole, our results support the idea that informant discrepancies result from meaningful behavioral variability and are more than measurement error. For readers familiar with debates in the personality literature (Mischel, 2009), this statement may create a sense of déjà vu (Mischel & Peake, 1982) all over again (Roberts & Caspi, 2001). Indeed, a critic might ask whether aspects of this history are repeating themselves in the informant discrepancy literature. Achenbach and colleagues (1987) argued in their influential meta-analysis that informant discrepancies result from children's behavioral variability across settings, yet, more than 20 years later, this message has not been fully absorbed. Researchers continue to use measures that are rooted in a nomothetic trait tradition (see Dumenci, Achenbach, & Windle, 2010) and that do not explicitly incorporate psychosocial contexts into the core of the measurement

process (see Block, 1995, 2010). To make progress, informant discrepancy researchers need to be alert to the notion, popularized in textbooks, that "consistency across situations lies at the core of the concept of personality" (Weiten, 2004, p. 478). Alternative, more contextualized concepts of personality do not require consistency in this sense, do not insist that complex patterns of behavior be aggregated across situations in order to arrive at a reliable index of a trait (Roberts & Caspi, 2001), and thus may be better able to incorporate the kinds of discrepancies found in the cross-informant literature.

A related suggestion is to note the assumptions on which our measurements are based. To paraphrase Hotelling, Bartky, Deming, Friedman, and Hoel (1948), we sometimes choose our measures as we say our prayers—because they are found in highly respected books written a long time ago. The measures used widely in child assessment have many strengths, but they do not come with a guarantee that they will measure only the "person" merely because they ask the rater about his or her "behavior." We should study not only the broad factors that predict syndromal measures or their differences across settings but also the narrow microcontexts that may be implicit in the mind of the rater when rating individual acts and that give meaning to their ratings and disagreements about them. We should also strive to assess the "environment" with the same rigor that we assess the "child" (e.g., Moos & Moos, 1990). All of this may appear at first to be a distraction from the immediate problem of understanding why parent and teacher assessments are so different, but genuine progress will require a deeper understanding of how intertwined behaviors and the surrounding social situation really are.

REFERENCES

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*, 94–98.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington: Research Center for Children, Youth, & Families, University of Vermont.

Barkley, R. A. (1988). Child behavior rating scales and checklists. In M. Rutter, A. H. Tuna & I. S. Lann (Eds.), *Assessment and diagnosis in child psychopathology* (pp. 113–155). New York: Guilford.

Beck, K. H., Hartos, J. L., & Simons-Morton, B. G. (2006). Relation of parent–teen agreement on restrictions to teen risky driving over 9 months. *American Journal of Health Behavior*, *30*, 533–543.

Block, J. (1995). A contrarion view of the five-factor approach to personality description. *Psychological Bulletin*, *117*, 187–215.

Block, J. (2010). The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, *21*(1), 2–25.

Borg, I., & Groenen, P. J. F. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Burke, J. D., Pardini, D. A., & Loeber, R. (2008). Reciprocal relationships between parenting behavior and disruptive psychopathology from childhood through adolescence. *Journal of Abnormal Child Psychology*, *36*, 679–692.

Buss, D., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90*, 105–126.

Cervone, D., Shadel, W. G., & Jencius, S. (2001). Social-cognitive theory of personality assessment. *Personality and Social Psychology Review*, *5*, 33–50.

De Los Reyes, A., Goodman, K. L., Kliewer, W., & Reid-Quińones, K. (2008). Whose depression relates to discrepancies? Testing relations between informant characteristics and informant discrepancies from both informants' perspectives. *Psychological Assessment*, *20*, 139–149.

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in children's disruptive behavior. *Journal of Abnormal Child Psychology*, *37*, 637–652.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483–509

Drabick, D. A. G., Gadow, K. D., & Loney, J. (2008). Co-occurring ODD and GAD symptom groups: Source-specific syndromes and cross-informant comorbidity. *Journal of Clinical Child & Adolescent Psychology*, *37*, 314–326.

Dumenci, L., Achenbach, T. M., & Windle, M. (2010). Measuring context-specific and cross-contextual components of hierarchical constructs. *Journal of Psychopathology and Behavioral Assessment*. Advance online publication. doi:10.1007/s10862-010-9187-4

Endler, N. S., Parker, J. D. A., Bagby, R. M., & Cox, B. J. (1991). Multi-dimensionality of state and trait anxiety: The factor structure of the Endler Multidimensional Anxiety Scales. *Journal of Personality and Social Psychology*, *60*, 919–926.

Feldt, L. S., & Kim, S. (2006). Testing the difference between two alpha coefficients with small samples of subjects and raters. *Educational and Psychological Measurement*, *66*, 589–600.

Ferdinand, R. F., van der Ende, J., & Verhulst, F. C. (2004). Parent–adolescent disagreement regarding psychopathology in adolescents from the general population as a risk factor for adverse outcome. *Journal of Abnormal Psychology*, *113*, 198–206.

Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2008). Integrating dispositions, signatures, and the interpersonal domain. *Journal of Personality and Social Psychology*, *94*, 531–545.

Granic, I., & Patterson, G. R. (2006). Toward a comprehensive model of antisocial development: A dynamic systems approach. *Psychological Review*, *113*(1), 101–131.

Grills, A. E., & Ollendick, T. H. (2003). Multiple informant agreement and the Anxiety Disorders Interview Schedule for Parents and Children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 30–40.

Guion, K., Mrug, S., & Windle, M. (2009) Predictive value of informant discrepancies in reports of parenting: Relations to early adolescents' adjustment. *Journal of Abnormal Child Psychology*, *37*, 17–30.

Hartley, A. G., Zakriski, A. L., Wright, J. C., & Parad, H. W. (2009, April). *Understanding sources of cross-informant agreement in the assessment of child psychopathology*. Poster presented at the 2009 Society for Research in Child Development Biennial, Denver, CO.

Haynes, S. N., Mumma, G. H., & Pinson, C. (2009). Idiographic assessment: Conceptual and psychometric foundations of individualized behavioral assessment. *Clinical Psychology Review*, *29*, 179–191.

Hotelling, H., Bartky, W., Deming, W. E., Friedman, M., & Hoel, P. (1948). The teaching of statistics. *Annals of Mathematical Statistics*, *19*, 95–115.

Hood, B., Power, T., & Hill, L. (2009). Children's appraisal of moderately stressful situations. *International Journal of Behavioral Development*, *33*, 167–177.

Kolko, D. J., & Kazdin, A. E. (1993). Emotional/behavioral problems in clinic and nonclinic children: Correspondence among child, parent, and teacher reports. *Journal of Child Psychology & Psychiatry*, *34*, 991–1006.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*, 1566–1577.

Maccoby, E. E. (1998). *The two sexes: Growing up apart, coming together*. Cambridge, MA: Harvard University Press.

Mischel, W. (2009). From personality and assessment (1968) to personality science, 2009. *Journal of Research in Personality*, *43*, 282–290.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review*, *89*, 730–755.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268.

Moos, R., & Moos, B. (1990). *Life stressors and social resources inventory preliminary manual*. Palo Alto, CA: Stanford University Medical Center.

Noordhof, A., Oldehinkel, A. J., Verhulst, F. C., & Ormel, J. (2008). Optimal use of multi-informant data on co-occurrence of internalizing and externalizing problems: The TRAILS study. *International Journal of Methods in Psychiatric Research*, *17*, 174–183.

Parker, D. A., & Wood, L. M. (2008). Personality and the coping process. In G. Boyle, G. Matthews & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol. 1: Personality theories and models* (pp. 506–519). Thousand Oaks, CA: Sage.

Roberts, B. W., & Caspi, A. (2001). Authors' response: Personality development and the person–situation debate: It's déjà vu all over again. *Psychological Inquiry*, *12*, 104.

Scotti, J. R., Morris, T. L., McNeil, C. B., & Hawkins, R. P. (1996). *DSM–IV* and disorders of childhood and adolescence: Can structural criteria be functional? *Journal of Consulting and Clinical Psychology*, *56*, 1177–1191.

Shoda, Y., Mischel, W., & Wright, J. C. (1993). Links between personality judgments and contextualized behavior patterns: Situation-behavior profiles of personality prototypes. *Social Cognition*, *11*, 399–429.

Smith, R. E., Shoda, Y., Cumming, S. P., & Smoll, F. L. (2009). Behavioral signatures at the ballpark: Intraindividual consistency of adults' situation-behavior patterns and their interpersonal consequences. *Journal of Research in Personality*, *43*, 187–195.

Van Mechelen, I. (2009). A royal road to understanding the mechanisms underlying person-in-context behavior. *Journal of Research in Personality*, *43*, 179–186.

Van Mechelen, I., & Kiers, H. A. L. (1999). Individual differences in anxiety responses to stressful situations: A three-mode component analysis model. *European Journal of Personality*, *13*, 409–428.

Vansteelandt, K., & Van Mechelen, I. (1998). Individual differences in situation-behavior profiles: A triple-typology model. *Journal of Personality and Social Psychology*, *75*, 751–765.

Vansteelandt, K., & Van Mechelen, I. (2006). Individual differences in anger and sadness: In pursuit of active situational features and psychological processes. *Journal of Personality*, *74*, 873–910.

Weiten, W. (2004). *Psychology: Themes and variations* (6th ed.). Belmont, CA: Wadsworth.

Wright, J. C., Lindgren, K. P., & Zakriski, A. L. (2001). Syndromal versus contextualized personality assessment: Differentiating environmental and dispositional determinants of boy's aggression. *Journal of Personality and Social Psychology*, *81*, 1176–1189.

Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology*, *53*, 1159–1177.

Wright, J. C., & Zakriski, A. L. (2003). When syndromal similarity obscures functional dissimilarity: Distinctive evoked environments of externalizing and mixed syndrome boys. *Journal of Consulting and Clinical Psychology*, *71*, 516–527.

Wright, J. C., & Zakriski, A. L. (2001). A contextual analysis of externalizing and mixed syndrome boys: When syndromal similarity obscures functional dissimilarity. *Journal of Consulting and Clinical Psychology*, *69*, 457–470.

Youngstrom, E., Loeber, R., & Southamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*, 1038–1050.

Zakriski, A. L., Wright, J. C., & Underwood, M. K. (2005). Gender similarities and differences in children's social behavior: Finding personality in contextualized patterns of adaptation. *Journal of Personality and Social Psychology*, *88*, 844–855.

Psychology
2013. Vol.4, No.1, 1-10
Published Online January 2013 in SciRes (http://www.scirp.org/journal/psych)
DOI:10.4236/psych.2013.41001

An Experimental Analysis of the Assessment and Perception of Behavior Change: How Summary Measures Influence Sensitivity to Change Processes

Anselma G. Hartley[1], Jack C. Wright[1], Audrey L. Zakriski[2], Anne N. Banducci[3]
[1]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, Providence, USA
[2]Department of Psychology, Connecticut College, New London, USA
[3]Department of Psychology, University of Maryland, College Park, USA
Email: Anselma_Hartley@brown.edu

Received October 6th, 2012; revised November 6th, 2012; accepted December 4th, 2012

A series of experiments examined how summary assessment measures influence people's ability to detect change in behavior over time and across situations. Two measures that are often used to assess child behavior (Teacher Report Form) and adult personality (Five Factor Inventory) were examined. Each instrument led perceivers to focus on the overall frequency of targets' behavior, even when targets differed both in how they reacted to social events and in how often they experienced those events in their interactions with others. Although people adopted an overall frequency perspective when using summary measures, they detected changes in events and targets' *if … then …* reactions to events when using alternative context-specific measures. The findings demonstrate how summary trait methods can shift perceivers' attention away from situational factors and thereby yield trait scores that are insensitive to context-specific but potentially important changes in targets' social behavior.

Keywords: Personality; Social Perception; Assessment; Behavior Change; Social Context

Introduction

A potential conflict exists between the way people think about personality and how researchers assess it. On the one hand, researchers often emphasize the breadth and stability of traits and therefore use personality measures that aggregate over variability that may occur over time and situations (Mischel, 2009; Watson, 2004). On the other hand, social cognition research suggests that people incorporate situational information into their personality impressions (Kammrath, Mendoza-Denton, & Mischel, 2005; Smith & Collins, 2009). Despite the widespread use of "summary" trait measures in both child and adult assessment, little research has explored how social perceivers use them under laboratory conditions in which the relevant inputs can be isolated and manipulated. The present research illustrates how such methods can deepen our understanding of how summary trait measures influence perceivers' sensitivity to personality change. In this paradigm, we create targets who show different patterns of change over time in their social environments and in how they responded to them. We examine the possibility that summary trait measures lead perceivers to focus on overall behavior rates and to de-emphasize contextual information they might otherwise use. We test the further implication that this emphasis on overall frequencies leads raters to report that target behavior is stable over time even when targets show clear changes in how they respond to specific social situations.

Summary approaches have a long tradition in child and adult assessment. On widely used child measures (e.g., Teacher Report Form or TRF, Achenbach & Rescorla, 2001), an adult typically rates how well brief statements describe the child. Many of these statements focus on the frequency of behaviors ("teases a lot," "threatens people"), some include trait adjectives ("stubborn"), and less often they refer to the context in which the behaviors occur ("disobedient at school", "defiant, talks back"). Popular "Big Five" measures used to assess adult personality (e.g., NEO-PI-R and the NEO-Five Factor Inventory or FFI, Costa & McCrae, 1992) also include behavior frequency statements (e.g., "seldom sad or depressed"), trait adjectives ("is a cheerful, high-spirited person"), and statements that explicitly refer to behavior in context ("if he doesn't like people, he lets them know it"). Although these child and adult measures vary in how their items were generated and how often they refer to contexts, they share an essential feature: Both aggregate into summary scales that do not reveal what these contexts are, how often they occur, or how responses to them may vary. Such measures thus focus on mean-level behavior tendencies, and do not reveal individual differences in how people respond to specific contexts (Cervone, 2005; Cervone, Shadel, & Jencius, 2001).

Alternative models incorporate context into personality assessment by examining *if … then …* links between events that occur in a person's social environment (e.g., *if* provoked) and their reactions to them (e.g., *then* hostile) (Vansteelandt & Van Mechelen, 1998; Wright & Mischel, 1987). Studies adopting such approaches have demonstrated that personality is revealed not simply through overall trait or behavior levels, but through an individual's contextualized patterning of trait-relevant behavior (Fournier, Moskowitz, & Zuroff, 2008; Hartley, Zakriski, & Wright, 2011; Hoffenaar & Hoeksma, 2002; Smith, Shoda, Cumming, & Smoll, 2009). A complementary line of "socially situated" cognition research proposes that context plays an

important role in social perception and judgment (Reeder, Monroe, & Pryor, 2008; Smith & Collins, 2009). Although early studies on the "fundamental attribution error" (Ross, 1977) argued that situational influences are often ignored, subsequent research found that people do incorporate contextual information into their personality judgments, but when and how they do so depends on several factors (Gilbert & Malone, 1995). For example, people have difficulty integrating situational influences into their dispositional judgments when the salience of the stimuli is low and cognitive load is high (Chun, Speigel, & Kruglanski, 2002). People's ability to process behavioral and situational information also depends on their statistical knowledge and investment in the target (Schaller, 1992), and on their affective state (Hunsinger, Isbel, & Clore, 2011).

Despite considerable field research using summary measures (Gresham et al., 2010; Terracciano, McCrae, & Costa, 2009), little work has examined how perceivers use them under controlled laboratory conditions. Social cognition research has used experimental methods to study people's use of situational information (Chun et al., 2002; Kammrath et al., 2005; Trope & Gaunt, 2000), yet this work has not examined how summary trait measures influence what people encode in their ratings. Some researchers have claimed that summary measures are implicitly contextualized by the respondent even when items lack explicit contextual cues (Tellegen, 1991; Wood & Roberts, 2006), and are therefore sensitive to reaction patterns (Denissen & Penke, 2008). For example, items that contain trait adjectives (e.g., "thoughtful and considerate", "is a cheerful, high spirited person") might lead the rater to infer the situations that are most relevant and to judge how the target reacts when those situations are encountered. However, we are unaware of an experimental test of this idea. Other researchers have speculated that summary methods lead people to rely on global representations lacking in specific time or setting cues (Schwarz & Oyserman, 2011). Support for this argument is found in studies showing that summary measures lead people to ignore conditional *if … then …* links between events and reactions and focus instead on overall act frequencies (Wright et al., 2001). In the present study, we test the idea that summary measures—including popular child behavior measures and adult five-factor measures—are designed to assess overall behaviors, do this well, but in doing so miss changes in how people respond to specific social situations.

We extended past work in several ways. First, rather than focusing on a single time point, we created targets that changed over time, both in how often they encountered events ("event rates") and in the conditional probability of their responses to them ("reaction rates"). In Studies 1-2ab, peer provocation and adult discipline were the focal events and aggression was the focal reaction, as these are relevant to child assessment (Dirks, Treat, & Weersing, 2007). This yielded two targets who showed "converging" changes in event rates and reaction rates (i.e., both decreased or both increased), and thus their overall rates of aggression increased or decreased. The two other targets showed "diverging" changes: One experienced an increase in aversive events, but became less likely to respond aggressively to them; the other experienced a decrease in aversive events, but became more likely to respond aggressively. These targets are especially interesting because they show opposite changes in event and reaction rates, yet show no change in overall aggression rates. If summary measures track only overall rates, as we have proposed, they should distinguish between targets whose overall rates differ, but fail to distinguish between targets who show opposite reaction change but constant overall behavior rates. If, on the other hand, these measures are implicitly contextualized as others have suggested, they should distinguish between targets whose reactions to events changed over time, even if their overall behavior rates did not.

Second, we used both child and adult targets, and we examined both popular measures for studying child behavior (TRF; Achenbach & Rescorla, 2001) and adult personality (NEO-FFI; Costa & McCrae, 1992). In each of our experiments, participants used the instrument to rate the target at the end of one period of observation, and then again at the end of a second period. Studies 1-2ab focused on aggressive behaviors of children that are relevant to the TRF, and Study 3 focused on (dis) agreeable behaviors of adults that are relevant to the agreeableness domain on the FFI. Guided by past theorizing and evidence (Schwarz & Oyserman, 2011; Wright et al., 2001), we hypothesized that relevant scales on the TRF (aggression) and FFI (agreeableness) would be sensitive to changes in targets' overall behavior rates, but insensitive to differences between the diverging targets whose reactions changed in opposite directions.

Third, we examined whether participants can detect changes in rates of eliciting events and changes in targets' conditional reactions to them, even if this is not evident when they use summary trait measures. Based on people's sensitivity to context at a single time point (Chun et al., 2002; Wright et al., 2001), we predicted that participants' open-ended descriptions of targets would refer not only to their overall behavior tendencies, but also to events targets encountered and their event-specific reactions. We further expected that participants would differentiate between the diverging targets when explicitly asked to estimate how often targets encountered events and the conditional probability of their reactions to those events. Because people can have difficulty judging conditional probabilities (see Fox & Levav, 2004), we examined how two response formats—a typical rating format (e.g., Vansteelandt & Van Mechelen, 1998) versus a frequency-count estimation format (Gigerenzer, 2008)—influenced their performance. Support for these hypotheses would indicate that widely used summary assessment methods divert people's attention away from situation-specific changes in behavior they otherwise notice and thereby yield ratings that reflect only targets' overall behavior frequencies.

Study 1

We first examined change over time. Using a 2 (event rate) × 2 (reaction rate) × 2 (phase) design, we manipulated whether a target child experienced an increase or decrease in the probability of aversive events ("event rates"), and an increase or decrease in the conditional probability of aggressive behavior when those events occurred ("reaction rates"). We hypothesized that the TRF is primarily sensitive to base-rates, and thus should be influenced by all factors that contribute to overall behavior (i.e., events and reactions), and not just by targets' reaction rates. Thus, the TRF should be unable to distinguish between the functionally diverging targets even though one showed an increase in aggressive reactions to aversive events and one showed a decrease.

Method

Participants

Forty-three undergraduates from the pool in an introductory psychology class participated at Brown University. Three were removed: two who completed materials out of order, and one who did not understand the instructions. This yielded a sample of 40 (20 M, 20 W, M_{age} = 19.2 years, SD = 1.17). All studies reported were approved by Brown University's Institutional Review Board.

Materials

The experimental stimuli were based on Wright et al. (2001), but described the target at two points. The target was identified as a fictitious 11-year-old boy ("Dan") in a residential summer program. Participants viewed 32 vignettes that described the target at the beginning of the summer (Phase 1) and 32 that described him 9 weeks later (Phase 2). Four targets were created. One encountered an increase in aversive events and showed an increase in aggressive reactions to those events (E+/R+) ("+" = increase). The second showed a decrease in both event rates and reaction rates (E−/R−) ("−" = decrease). The third encountered an increase in aversive events, but showed a decrease in aggressive reactions (E+/R−). The fourth had the reverse arrangement (E−/R+).

Each vignette, presented for 9 seconds on an otherwise blank computer screen, described the setting and an interaction between Dan and another person. The setting, agent, agent action, target name, and response appeared in the same order. Events consisted of aversive peer events (tease, threaten), aversive adult events (warn, discipline), nonaversive peer events (prosocial talk, ask), and non-aversive adult events (prosocial talk, ask/instruct). Reactions were aggressive or nonaggressive. An example of a peer aversive event with an aggressive reaction is: "In the dining hall a boy says, 'Shut up and give me your dessert.' Dan replies, 'No, you shut up. I want it.'" An example of an adult aversive event with a non-aggressive reaction is: "In swimming, a counselor says, 'You better not go past that green rope.' Dan says, 'Okay, I won't.'"

Table 1 shows the probabilities of aversive events, p(E), the conditional probabilities of aggressive reactions to those events, p(R|E), and the corresponding frequencies. The probabilities of aversive events are obtained by dividing the number of aversive events per phase by the total number of vignettes per phase (32). Conditional probabilities of aggressive reactions are obtained by dividing the number of aggressive behaviors to aversive events by the number of aversive events encountered. The overall probability or "base rate" of aggressive behaviors, p(R) is obtained by p(E)∗p(R|E); this is equivalent to the number of aggressive behaviors per phase divided by the total number of vignettes per phase. The converging E+/R+ and E−/R− targets showed increases (or decreases) both in aversive events and in aggressive reactions to them, and therefore their base rates of aggression increased (or decreased) over phases. The diverging E−/R+ and E+/R− targets (rows 2 - 3) differed in the conditional probability of their aggressive reactions to aversive events, but had equal base rates of aggression at each phase.

Table 1.
Properties of the four experimental targets for all studies.

Condition	Phase 1			Phase 2		
	p(E)	p(R|E)	p(R)	p(E)	p(R|E)	p(R)
E−/R−	.75	.75	.56	.25	.25	.06
	(24/32)	(18/24)	(18/32)	(8/32)	(2/8)	(2/32)
E−/R+	.75	.25	.19	.25	.75	.19
	(24/32)	(6/24)	(6/32)	(8/32)	(6/8)	(6/32)
E+/R−	.25	.75	.19	.75	.25	.19
	(8/32)	(6/8)	(6/32)	(24/32)	(6/24)	(6/32)
E+/R+	.25	.25	.06	.75	.75	.56
	(8/32)	(2/8)	(2/32)	(24/32)	(18/24)	(18/32)

Note: p(E) = probability of aversive event; p(R|E) = probability of aggressive reaction to aversive event; p(R) = base-rate probability of aggressive behavior. Note that p(R) = p(E) ∗ p(R|E). "+" indicates increase; "−" indicates decrease in event or reaction rate. E = event; R = reaction. Values in parentheses indicate frequencies on which probabilities and conditional probabilities were based; for p(E) and p(R), the denominator is always the total number of vignettes per phase (32), and for p(R|E), the denominator is the number of aversive events per phase.

Dependent Measures

Open-Ended Descriptions. Participants read, "You've just read about Dan during the first week of June (second week of August) in the residential summer program. Please describe in a few sentences what was most important about Dan and the summer program during that time."

Teacher Report Form. As in Wright et al. (2001), we used a subset of the 118 items from the 1993 version of the TRF (Achenbach, 1993) to avoid fatigue. Specifically, we used the scale that was most relevant to this study (aggression, 25 items) and a contrast scale (withdrawal, 9 items), with "school" changed to "camp" for our stimuli. An example of an aggression item is "argues a lot"; an example of a withdrawal item is "unhappy, sad, or depressed." Items were rated using the TRF's 0 - 2 scale. Test-retest reliability of the TRF aggression and withdrawal scales in field studies is reported to be .89 and .85 respectively when the interval is 2 - 3 weeks (Achenbach, Howell, McConaughy, & Stanger, 1995). The TRF aggression scale correlates modestly but significantly with classroom observations of verbal aggression and disruptive behavior (Henry, 2006).

Perceived Overall Change. Participants rated changes in Dan's "overall behavior", "behavior toward peers", and "behavior toward counselors". These were averaged into an "overall target change" scale (α = .96). Next, they rated how peers' and adults' overall "behaviors towards Dan changed." These were averaged into an "overall social environment change" scale (α = .96). All items used a 7-point scale (1 = much worse, 7 = much improved).

Behavior, *Event*, *and Reaction Measures.* To clarify whether participants detected overall behavior rates, event rates, and reaction rates at each phase, these items corresponded as closely as possible to the stimuli. Participants first rated the overall frequency of the target's aggressive and prosocial behaviors shown during Phase 1 using 4 items (e.g., "Dan argued or quarreled", "talked politely/made friendly requests"). They then rated how often Dan encountered aversive and non-aversive events at Phase 1, using 4 items (e.g., "peers teased, threa-

tened, or bossed Dan", "adults complimented/made friendly requests"). Next, they rated the target's reactions given that some event occurred, using 16 items (4 events × 4 reactions). Participants read, "Indicate how often Dan showed each reaction to the event described." After each of 4 event prompts ("If a peer teased, threatened, or bossed Dan …"), the participant rated how often the target showed a reaction to it (e.g., "he argued or quarreled"); the wording of the reaction was the same as the wording of the behaviors noted above. Participants then rated the behaviors, events, and reactions that were shown during Phase 2. All items were rated on a 6-point scale (0 = never, 5 = almost always).

Procedure

Participants were run in groups of 1-4 on separate computers and were randomly assigned to condition, to which the experimenter was blind. Using the dependent measures just described, participants completed these steps, in order: 1) read 32 vignettes for Phase 1, each for 9 s; 2) open-ended description and TRF; 3) 32 vignettes for Phase 2; 4) repeat step 2; 5) overall perceived change; 6) additional ratings of behavior, events, and reactions seen at Phase 1 and at Phase 2. To avoid contaminating the TRF, it was administered before measures that mentioned events or reactions.

Preliminary Analyses

Participants' open-ended responses were coded as follows. 1) "Overall behavior": An uncontextualized statement about a prosocial, neutral, or aggressive behavior or disposition without a specified eliciting event (e.g., "Dan was friendly"). 2) "Event": A statement about a positive, neutral, or aversive event without a specified response (e.g., "People were nice to Dan"). 3) "Reaction": A prosocial, neutral, or aggressive behavior in response to a positive, neutral, or aversive event (e.g., "Dan was friendly when others were nice to him"). Agreement between the first author and a coder who was blind to condition was acceptable (average $\kappa = .80$).

Additional analyses examined how perceived overall change measures (see previous) compared with other measures. The perceived overall change scale correlated highly with the calculated TRF aggression change ($r = .88$, $p < .001$), and the perceived overall social environment change scale correlated highly with the calculated event change score ($r = .93$, $p < .001$). To avoid redundancy, perceived overall change analyses are not presented.

Results and Discussion

Open-Ended Descriptions

Although the open-ended descriptions were not our main focus, we examined the Phase 1 descriptions to clarify participants' perceptions before they were affected by the TRF. Based on past research (Kammrath et al., 2005), we predicted that participants would not only describe overall behavior tendencies, but also describe events and conditional reactions to them. We calculated percentages by dividing the number of statements in each category for each participant by the total number of codeable statements for that participant. As predicted, participants used all statement types, with nonsignificant differences in their mean relative frequency: uncontextualized behavior statements (40%), event statements (32%), and reaction statements (28%), $F(2, 72) = 2.15$, $p > .1$. We also found a statement type x reaction condition interaction, $F(2, 72) = 6.18$, $p < .005$, $\eta^2 = .15$. In conditions with low reaction rates at Phase 1, uncontextualized behavior statements were more frequent (52%) than event statements (26%) or reaction statements (22%), whereas in conditions with high reaction rates at Phase 1, statement types differed less (28%, 38%, and 34%, respectively). We found a similar pattern when analyses were restricted to statements about aggressive behaviors; details can be obtained from the first author.

Summary Trait Assessment

We expected that the TRF would detect changes in overall behavior rates, but not distinguish between the functionally diverging targets whose overall rates were equal. Specifically, we predicted that TRF aggression ratings would decrease over phase for the E−/R− condition, increase for the E+/R+ condition, and remain unchanged for the diverging conditions (E−/R+, E+/R−).

As shown in **Figure 1**, the results supported this prediction. A 2 (event) × 2 (reaction) × 2 (phase) ANOVA, with phase as a repeated measure, revealed the expected reaction condition x phase interaction, $F(1, 36) = 56.99$, $p < .001$, $\eta^2 = .61$. Also as expected, we found an interaction between event condition and phase, $F(1, 36) = 7.24$, $\eta^2 = .66$. (In all repeated-measures analyses, significance tests were based on Greenhouse-Geisser adjustments.) We also found a small unexpected effect for phase, $F(1, 36) = 5.52$, $p < .05$, $\eta^2 = .13$; TRF aggression ratings were slightly higher overall at Phase 1 than Phase 2. No other effects were expected or found.

To simplify subsequent analyses, we computed change scores (Phase 2 - Phase 1), which were then submitted to a 2 (event condition) × 2 (reaction condition) ANOVA. **Figure 2(A)** presents mean TRF change in standardized form (*z*-scores); this was solely to permit graphical comparisons with other measures with different natural metrics, and otherwise had no effect on any findings we report. Our predictions and findings necessaryily parallel those just explained, though are now expressed as change scores. We found the expected main effects for event and reaction condition (**Table 2**) and the expected Tukey's HSD comparisons (**Figure 2(A)**). As predicted, the TRF was sensitive to changes in overall behavior, but not to the event or reaction changes that contributed to those rates. As shown in **Figure 2(A)**, the diverging conditions (E−/R+, E+/R−; see middle bars) with identical overall behavior rates in the stimuli did not differ for TRF aggression despite the fact that one increased in aggressive reactions and the other decreased.

The preceding analyses used categorical predictors (condition), and do not fully reveal how participants' ratings were predicted by the base-rates of aggressive acts in the stimuli. Recall that values for $p(R)$ can be derived by multiplying $p(E)$ and $p(R|E)$ as shown in **Table 1**. Because this (equal) weighting yields the base rates, we expected it to best predict the TRF aggression ratings. It is also possible that participants were more influenced by the probability of encountering events, or by the conditional probability of reactions to them. To test this, we attached weights between .01 - .99 (in increments of .01) to each component and computed predicted values. With w as the event weight, and $1 - w$ for the reaction weight, the predicted values were $[(w_i p(E) + (1 - w_i)p(R|E)]/2$. For each weighted set,

Table 2.
F-tests and effect sizes for ANOVAs of Teacher Report Form (TRF) ratings, event judgments, and reaction judgments, for Studies 1-2ab.

		TRF		Event		Reaction	
Study	Source	*F*	η^2	*F*	η^2	*F*	η^2
1	Reaction	56.99	.61	10.77	.23	126.54	.78
	Event	70.24	.66	137.38	.79	42.42	.54
	R × E	.32	.01	1.56	.04	1.85	.05
2a	Reaction	40.90	.53	12.46	.26	92.89	.72
	Event	47.02	.57	154.74	.81	25.85	.42
	R × E	2.39	.06	8.17	.19	1.19	.03
2b	Reaction	90.75	.72	8.87	.20	50.78	.59
	Event	94.78	.73	45.25	.56	.95	.02
	R × E	.03	.00	.02	.00	.08	.00

Note: R × E = Reaction × Event interaction. Degrees of freedom were (1, 36) for all studies. All *F*'s > 7.40 (12.83) were significant at $p < .01$ (.001); all other *F*'s shown were $p > .05$.

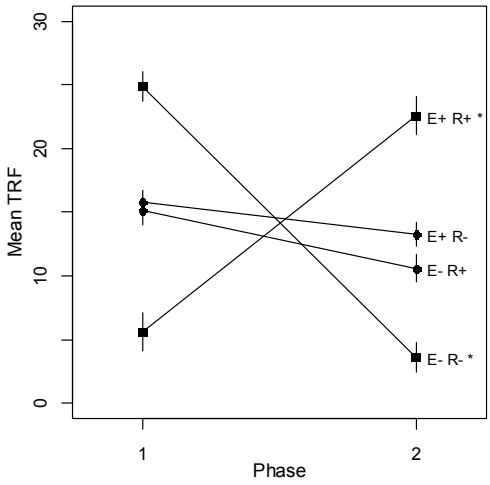

Figure 1.
Mean Teacher Report Form (TRF) aggression ratings by phase, for Study 1. Experimental conditions are shown next to each line. Error bars indicate +/− 1 SEM. Asterisks indicate significant differences across phase (*p*s < .001).

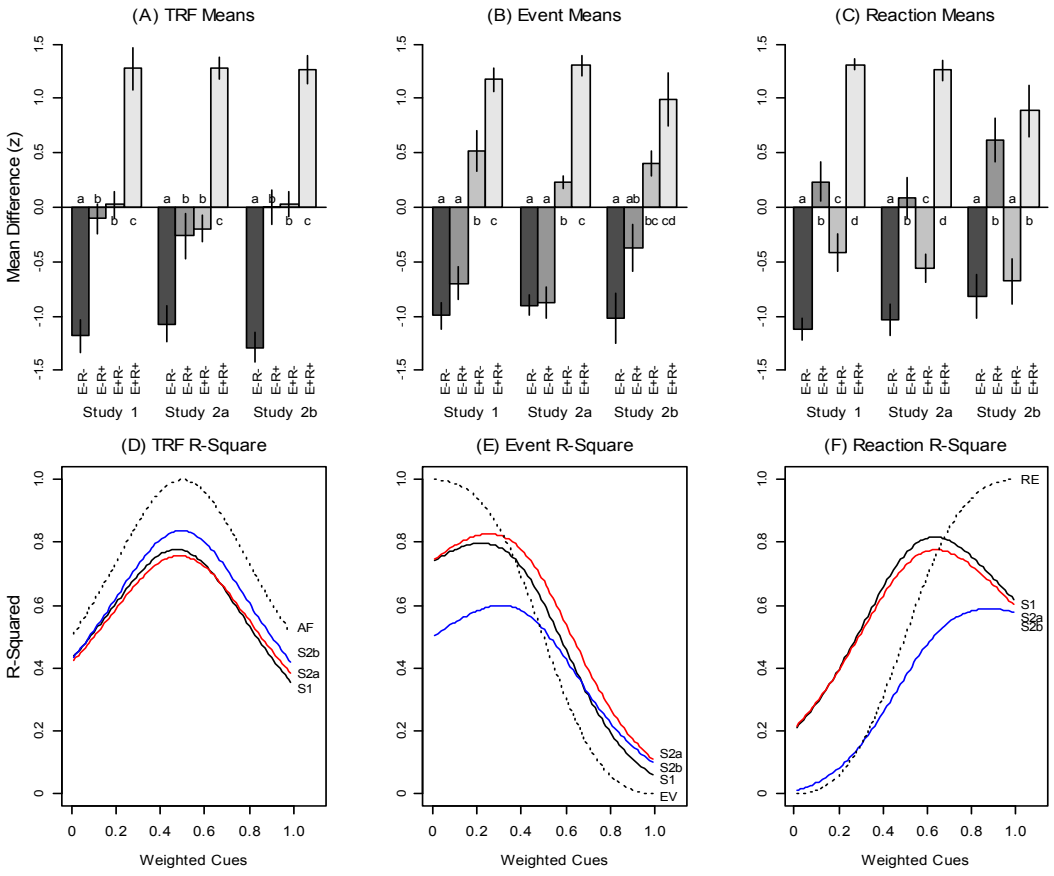

Figure 2.
Results for Teacher Report Form (TRF), event, and reaction measures for Studies 1 (S1), 2a (S2a), and 2b (S2b). Top row (panels A-C) shows mean change scores for each measure (standardized within study). Experimental conditions are on the abscissa. Bars within a panel that do not share a subscript (a)-(d) are significantly different based on Tukey's HSD. Error bars indicate +/− 1 SEM. Bottom row (panels (D)-(F)) shows cue weight analysis results for TRF, event, and reaction judgments, respectively. A "weighted cue" value of 0 on the abscissa represents a full weighting of events; 1 represents a full weighting of reactions. The ordinate shows the R^2 values for predictions of participants' ratings for phases 1 and 2 combined. Dotted lines indicate hypothetical perfect sensitivity to act-frequencies (AF); events (EV), and reactions (RE).

we computed scores from these values, used them to predict participants' deviation from their mean TRF aggression rating over the two phases, and computed R^2. If participants showed perfect sensitivity to the base rate of aggression, a peak R^2 of 1.0 would occur at equal weighting of events and reactions (.50 on the abscissa; see line "AF" in **Figure 2(D)**). Perfect sensitivity to events is shown by line "EV" in **Figure 2(E)**; perfect sensitivity to reactions is shown by line "RE" in **Figure 2(F)**. As expected, results for the TRF resembled the theoretically perfect AF curve in **Figure 2(D)** (see "S1" for Study 1), and were best modeled (R^2 = .81) when event rates (.55) and reaction rates (.45) were nearly equally weighted.

Event Judgments

We examined participants' judgments of events using the same method as for the TRF. We predicted that event judgments would show increases in the E+ conditions and decreases in the E− conditions. As expected, the largest effect was the main effect for event condition (**Table 2**), with judged event change higher on average for the E+ conditions and pairwise comparisons showing discrimination between the functionally diverging conditions (**Figure 2(B)**). We also found a smaller, unexpected main effect for reaction condition, with judged event change higher on average for R+ conditions. As shown in **Figure 2(B)**, the mean change for the E+/R− condition, though in the expected direction, was lower than one would expect if participants' event ratings were influenced only by events. As shown in **Figure 2(E)**, results for participants' event judgments resembled the theoretical results (see line "EV") and were best modeled (R^2 = .80) when the weight was high for event rates (w = .78) and low for reaction rates (.22).

Reaction Judgments

Parallel analyses were performed for judgments of aggressive reactions to aversive events. We expected participants to be sensitive to changes in target's reaction rates and for their ratings to increase in the R+ conditions and decrease in the R− conditions. As expected, the largest effect was the main effect for reaction condition (see **Table 2**), with pairwise comparisons showing discrimination between the diverging conditions (**Figure 2(C)**). However, we also found a main effect for event condition; the marginal mean was higher for E+ conditions. As shown in **Figure 2(C)**, the mean changes for the diverging conditions (E−/R+, E+/R−), were not as large as one would expect if reaction ratings were influenced only by reaction rates. As shown in **Figure 2(F)**, reaction ratings were best modeled (R^2 = .82) when the weights were less extreme (w = .63 for reactions, .37 for events) than was found for event judgments. Compared to the results for event judgments, these results do not correspond as closely to the theoretically perfect results (see line "RE").

Summary

As expected, the TRF aggression scale was sensitive to changes in the overall rate of targets' aggression. It did not detect differences between targets whose base rates were unchanged, even though one of them increased in aggressive reactions and the other decreased. Although participants focused on act frequencies when using the TRF, they detected how often targets encountered events and their conditional reactions to those events when context-sensitive measures were used. This occurred even though they provided these judgments at the end of the experiment, when memory demands were high. Participants' reaction judgments were influenced more than anticipated by how often the targets encountered relevant events.

Studies 2a-b

One interpretation of participants' relative difficulty judging reaction rates is that the changes they observed violated their expectations about the stability of behavior over time. For example, some studies suggest that temporal stability is high relative to the cross-situational consistency of behavior (Fleeson, 2001), and that people over-rely on the former when making judgments about personality (Mischel & Peake, 1982). Study 2a therefore examined whether participants' judgments would be more sensitive to reaction changes when targets' behavior varied across settings (i.e., classrooms) rather than over time as in Study 1. A second interpretation is that judging reactions to events is more complex than judging overall behavior rates or event rates. Past research demonstrates that people have difficulty interpreting conditional probabilities (Fox & Levav, 2004) and that formally equivalent tasks may be easier when they are presented in a frequency-count format (Gigerenzer, 2008). To address these questions, Study 2b reformatted the event and reaction dependent measures into a frequency-count format and asked participants to provide separate estimates of how often events and relevant reactions to those events occurred.

Method

Participants. For Study 2a, 40 students (23 W, 17 M, M_{age} = 21.22 years, SD = 3.50) participated, and for Study 2b, 40 (21 W, 19 M, M_{age} = 22.92 years, SD = 3.82) participated. Participants in both studies were recruited from the Brown University community through flyers advertising a "psychology study" and were paid $8 for volunteering.

Materials and procedure. For Studies 2a-b, stimuli were nearly identical to those in Study 1, but minor revisions were made to describe cross-situational change rather than temporal change. Whereas Study 1 described the target's behavior at two distinct points in time (June and August) Studies 2a-b described the target's behavior in two classroom settings (art and music). Otherwise, the specific events and reactions described were the same as those used in Study 1.

Study 1 used items from the 1993 TRF to determine if the findings from Wright et al.'s (2001) study of behavior at a single time point extended to behavior change. Study 2a-b used items from the 2001 TRF to determine if our results generalize to the more recent version of the instrument. The aggression scales in the two versions are similar, with 19 of the 20 items in the 2001 version also appearing in the 1993 version (see Achenbach & Rescorla, 2001). The remaining dependent measures in Study 2a were identical to those used in Study 1, with minor word changes to ask about cross-situational change. For example, when participants were asked about the target's behavior at Phase 1, the word "June" was changed to "art class"; likewise for Phase 2, "August" was changed to "music class." Study 2b was identical to Study 2a, except that the behavior, event, and reaction measures were changed from a rating format (see Study 1, Method) into a frequency-count format. Par-

ticipants were first asked to report the overall frequency of the target's behaviors, or *n*(R), at Phase 1 and Phase 2. The program required that participants' answers be between 0 - 32. The same format was used for event judgments, *n*(E). Using the *n*(E) estimate provided, the reaction prompt read, "You reported that peers teased Dan [*n*(E)] times. Out of those [*n*(E)] times, how many times did Dan respond by arguing or quarreling?"; we refer to this as *n*(R ∩ E), where ∩ = the intersection of reactions and events. Answers were required to be between 0 and *n*(E) previously estimated. We computed the conditional probability of a reaction given an event ("computed reaction") as, *p*(R|E) = *n*(R ∩ E)/*n*(E).

Results and Discussion

As predicted, the results for TRF ratings for Studies 2a and 2b were similar to those in Study 1 and again supported the hypothesis that the TRF would be sensitive to overall behavior rates, and not detect changes in diverging targets. The main effects (**Table 2**), pairwise comparisons (**Figure 2(A)**), and cue weighting analyses (**Figure 2(D)**) were similar to those for Study 1. As expected, TRF ratings for Studies 2a-b were best predicted (R^2 = .77 and .82, respectively) when weights for events (*w* = .50) and reactions (.50) were equal, as would occur for ideal act frequency sensitivity.

The results for Study 2a again supported the hypothesis that participants would be sensitive to changes in events. The expected main effect for event condition was obtained, as was a smaller effect for reaction condition (**Table 2**). As expected, participants detected the difference between the events rates for the E+/R− and E−/R+ targets, but again they were also somewhat affected by reaction rates. Participants' event ratings (**Figure 2(E)**) were best predicted (R^2 = .83) when the weight was high for events rates (*w* = .75) and low for reaction rates (.25), as expected.

For reaction judgments, the expected main effect for reaction condition was found, as was the now familiar, smaller main effect for event condition (**Table 2**). Change for the diverging conditions (E−/R+, E+/R−) was differentiated (**Figure 2(C)**), but less clearly than one would expect if reaction ratings were solely influenced by reaction rates. Reaction judgments (**Figure 2(F)**) were best predicted (*r* = .78) when the weight was higher for reaction rates (*w* = .64) than for event rates (.36). Thus, the results essentially replicated those in Study 1; the cross-setting format of Study 2a did not measurably affect participants' sensitivity to reaction change.

Although the cross-setting format did not seem to increase participants' sensitivity to reaction change, we expected the frequency-count format used in Study 2b to increase participants' sensitivity to event rates and reaction rates by decoupling the conditional probability format of the reaction rating task. For event judgments, we found the expected main effect for event condition (**Table 2**), and change scores for the diverging conditions (E−/R+, E+/R−) were in the expected direction (**Figure 2(B)**). However, mean change was less extreme than expected for both diverging conditions (E+/R−; E−/R+), and participants demonstrated slightly *less* sensitivity to events using this response format. Compared to Studies 1-2a, event judgments were predicted (R^2 = .60) by a weighted combination of events (*w* = .65) and reactions (.35) (see **Figure 2(E)**).

In contrast, the frequency-count format did increase participants' sensitivity to reaction change. The computed conditional probabilities were uniquely influenced by the actual conditional probabilities of targets' reactions (**Table 2**). As shown in **Figure 2(C)**, the means for the diverging conditions (E−/R+, E+/R−) were different and now comparable to the converging conditions with corresponding reaction change (E+/R+, E−/R−). The cue weight analysis (**Figure 2(F)**) showed that the reaction measure was best predicted when the reaction weight was relatively high (*w* = .88) and the event weight was low (.12). However, **Figure 2(F)** also reveals that the means in the converging conditions were less extreme and the reaction measures more variable (i.e., standard errors larger) than in previous studies, resulting in a lower peak R^2 value (.59).

Summary. As in Study 1, in Studies 2a-b, TRF ratings were predicted by the actual base-rates of aggressive acts, and did not distinguish between targets who showed equal overall change, but opposite changes in aggressive reactions. As in Study 1, participants' event judgments were sensitive to actual event rates, though they were somewhat influenced by reaction rates. For Study 2b, event judgments were influenced by actual event rates, but were noisier when the frequency-count format was used. In contrast, the frequency-count format in Study 2b improved participants' sensitivity to reaction change: Conditional probabilities derived from participants' frequency estimates were influenced solely by changes in the conditional probabilities of targets' reactions. These results indicate that people can assess change in reactions but have some difficulty under the conditions we created, and improve when the frequency-count format is used.

Study 3

One might argue that our findings for the child assessment method (TRF) do not apply to widely-used adult personality measures (e.g., NEO-FFI; Costa & McCrae, 1992). As we have noted, some researchers have argued that five-factor measures may emphasize behavior frequencies less and allow observers to give greater weight to targets' conditional reactions (see Wood & Roberts, 2006) and therefore detect reaction patterns (Denissen & Penke, 2008). If so, the FFI could distinguish between our functionally diverging, but act-frequency equivalent targets. We suggest, however, that the majority of the FFI's items are act frequency in nature, and we therefore predicted that the FFI, like the TRF, would be primarily affected by changes in the frequency of targets' trait-relevant behaviors. Study 3 therefore focused on the FFI domain of agreeableness and created stimuli that were structurally identical to those used in Studies 1-2ab, but described a college student showing (dis)agreeable reactions to (non)aversive events. Although agreeableness (A) was the main interest, all domains were analyzed. We expected other domains that were relevant to our stimuli—extraversion (E) and neuroticism (N)—to behave similarly to agreeableness, and not distinguish between functionally diverging targets. We made no predictions for openness (O) and conscientiousness (C), as these behaviors were not the focus of the study.

Method

Thirty-nine undergraduates (23 W, 16 M, M_{age} = 19.21 years, *SD* = 1.10) from an introductory psychology pool participated. Stimuli had the same event and reaction rates as in Study 1, but described a 19 year-old sophomore, and focused on agreeable-

ness. Because the target was an adult, interactions involved only peers (rather than peers and adults). An example of an aversive event paired with a disagreeable reaction is: "Dan's lab partner says, 'I don't want to do the analyses in the way we agreed.' Dan replies, 'Tough. We're doing it my way and I'm not changing my mind.'" The dependent measure was the 60-item NEO-FFI (Costa & McCrae, 1992).

Results and Discussion

FFI scale scores were primarily sensitive to changes in act frequencies. As shown in **Figure 3(A)**, the three traits most relevant to the experiment (A, E, N) showed results that were similar to those for the TRF in Studies 1 and 2. There were main effects for reaction condition, F's(1, 35) > 2.56, ps < .001, η^2's = .37 (N), .54 (E), and .74 (A), main effects for event condition F's(1, 35) > 39.36, ps < .001, η^2's = .53 (N), .61 (E), .63 (A), and no significant interactions nor discrimination between functionally diverging targets. As predicted, participants' A, E, and N ratings were best predicted by a weighted combination of events (.45, .54, .59, respectively) and reactions (.55, .46, .41) (**Figure 3(B)**), which were all similar to the ideal act frequency result. For O, there was a main effect for reaction condition, F(1, 35) = 19.86, p < .001, η^2 = .36, and for C a main effect for event condition, F(1, 35) = 15.01, p < .001, η^2 = .3. Although the R^2 values for O and C were lower than for the other traits, O ratings were better predicted by reactions (.61) than by events (.39), whereas the C ratings were better predicted by events (.75) than by reactions (.25).

General Discussion

This research used an experimental approach to examine the perception and assessment of behavior change. Three main findings emerged. First, two instruments that are widely used in child and adult assessment enabled raters to detect changes in overall behavioral tendencies, but did not enable them to distinguish between targets who showed opposite changes in their trait-relevant reactions to events. Second, in both temporal (Study 1) and cross-situational paradigms (Study 2a), participants were sensitive to changes in the social events the target encountered. Third, participants were sensitive, but somewhat less so, to the conditional probability of targets' reactions to those events when explicitly asked to assess them. These results support the view that popular child and adult summary measures assess overall behaviors rather than reactions. They also demonstrate that such measures can show stability even when changes occur in people's reactions to events, and illustrate how people's perceptions of change may diverge from conclusions based on their own summary trait ratings.

We have noted that people might "implicitly contextualize" items on child behavior checklists and adult personality inventories, even though most items in such measures do not explicitly identify the context in which a behavior may occur (see Denissen & Penke, 2008; Tellegen, 1991; Wood & Roberts, 2006). In this view, the rater infers the situations that are most relevant and focuses on the target's conditional responses to those situations. We predicted, however, that these measures would primarily assess overall behaviors and show little sensitivity to people's reaction patterns. Our results supported this prediction and provided little evidence of implicit contextualization for either of the measures we studied. The aggression scale on the child measure (TRF) distinguished between the targets based on their overall behavior frequencies. However, it did not distinguish between targets who showed opposite patterns of change in their social environments and how they reacted to them. Likewise, domain scores on the adult measure (FFI) also appeared to be primarily sensitive to overall behavior and did not distinguish between changes that originated in the environment versus those that originated in the target's reactions.

The summary instruments we examined were built on the assumption that personality is stable and enduring, and therefore focus on mean-level behaviors rather than situational influences (see Cervone et al., 2001). In this regard, our results show that the TRF and FFI capture precisely what they were designed to capture: overall behavior. However, our results also highlight the tradeoffs associated with this emphasis on overall from changes in the social situations they encounter. Our studies

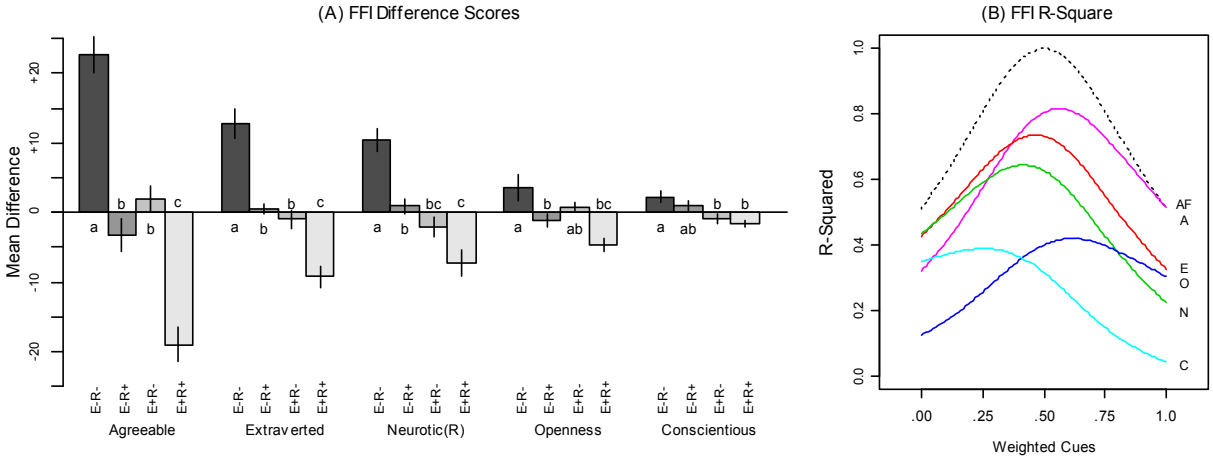

Figure 3.
Results for NEO-FFI for Study 3. Panel A shows mean change scores for agreeableness (A), extraversion (E), neuroticism (N), openness (O), and conscientiousness (C). Experimental conditions are on the abscissa. Bars within a panel that do not share a subscript (a)-(c) are significantly different based on Tukey's HSD. Error bars = +/− 1 SEM. Panel B shows cue weight analysis for FFI judgments for A, E, N, O, and C. AF = hypothetical perfect sensitivity to act-frequencies

also illustrate how summary measures could show that behavior is stable over time or across settings even when an individual shows clear changes in how they respond to social stimuli. These findings suggest that research on change over time and across settings (see Helson, Jones, & Kwan, 2002, Terracciano et al., 2009) should not over-rely on summary trait or behavior measures, but should also incorporate measures that explicitly examine people's reaction patterns and the make-up of their social environments.

Overall, our findings from the event and reaction rating tasks indicate that, given the right assessment format, participants can report on events and reactions when asked. However, they also indicated that judgments about reactions, $p(R|E)$, may be inherently more difficult than overall frequency judgments because they require the perceiver to encode how often an event occurred as well as how often a behavior co-occurred with it. We attempted to improve participants' performance in Study 2b by decomposing the task into its two frequency components: participants first estimated the frequency of aversive events, $n(E)$, and then estimated the frequency of aggressive acts to those events, $n(R \cap E)$. We then computed conditional probabilities from these two estimates in the usual fashion, $p(R|E) = n(R \cap E)/n(E)$. These derived estimates were affected uniquely by the actual conditional probabilities of targets' reactions in the stimuli, and were not influenced by how often targets encountered events, as found in Study 1 and 2a. A key challenge for future research is to determine the task formats that best enable people to disentangle event rates and reaction rates, but that are as simple and efficient as possible.

Interpreting participants' difficulty in judging reactions requires careful attention to our procedure. The reaction measure in Studies 1-2ab was administered for both Phase 1 and 2 after participants had filled out the TRFs. Completing the act frequency task first may have framed all subsequent measures in the experiment and may have influenced participants to think more as "act frequentists" rather than "contextualists" (see Schwarz & Oyserman, 2011; Wright et al., 2001). Findings from the open-ended assessments provide some support for this interpretation. Participants' initial descriptions of the targets, which were provided before they were influenced by other measures at Phase 1, not only used uncontextualized behavior statements, but also used simple event statements and conditional *if … then …* statements about event-reaction links.

Limitations of our studies should be noted. First, although our experimental approach answers questions about how summary assessments measure change, our manipulations for the event and reaction change parameters were larger (.25/.75) than might typically be observed in natural settings. Additional laboratory studies will be needed to examine how the TRF, FFI, and other summary measures (e.g., BFI; John, Donahue, & Kentle, 1991) perform under a wider range of stimulus manipulations. It will also be important to examine measures that appear to give greater emphasis to children's reactions to events (e.g., SSRS, Gresham & Elliot, 1990) and those that also focus on features of the social environment (e.g., Fournier et al., 2008).

Second, because our focus was on the TRF and FFI, other measures were either brief (e.g., open-ended descriptions) or were collected after all stimuli were shown. In contrast to other research on people's use of contextual information (Chun et al., 2002; Wright et al., 2001), our studies required subjects to encode multiple interactions over two phases, and only then estimate events and conditional reactions at Phases 1 and 2. This put the retrospective event and reaction ratings at a disadvantage. However, field studies often involve even more challenging conditions, in which raters' are asked to summarize more complex social interactions over much longer time periods. Clearly, additional research will be needed to answer questions about how people use information about situations and reactions under a wide range of stimulus complexity and memory load conditions (see Chun et al., 2002).

Overall, our findings suggest that instruments widely used to study personality change research are efficient at assessing overall behavior change, but ill-equipped to capture nuanced, context-specific dispositional and environmental change processes. As a result, these measures may have difficulty revealing whether behavior change stems from changes in the person, the environment, or both. Given our findings that people are sensitive to changes in the environment and in people's reactions (given the proper assessment format), it should be possible to develop measures that are more consistent with how people naturally encode behavior in context and that are better suited to assess the context-specific aspects of personality change. A major goal of future research in this area should be to deepen our understanding of the judgment processes that are engaged (or disengaged) when informants complete an assessment instrument, and use that knowledge to help improve the quality of assessment practices in research and applied settings.

Acknowledgements

This research was supported in part by award number R15MH076787 and 3R15MH076787-01S1from the National Institute of Mental Health. The content is solely the responsebility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. We are especially grateful to David Freestone, whose programming assistance made it possible to collect the data reported in Study 2b. We also thank Russell Church and Elena Festa Martino for their comments on earlier versions of this work.

REFERENCES

Achenbach, T. M. (1993). *Empirically based taxonomy: How to use syndromes and profile types derived from the CBCL/4-18, TRF, & YSR*. Burlington: University of Vermont.

Achenbach, T. M., Howell, C. T., McConaughy, S. H., & Stanger, C. (1995). Six-year predictors of problems in a national sample of children and youth: I. Cross-informant syndromes. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*, 336-347.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont. doi:10.1097/00004583-199503000-00020

Cervone, D., Shadel, W. G., & Jencius, S. (2001). Social-cognitive theory of personality assessment. *Personality and Social Psychology Review, 5,* 33-50. doi:10.1207/S15327957PSPR0501_3

Cervone, D. (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology, 56,* 423-452. doi:10.1146/annurev.psych.56.091103.070133

Chun, W. Y., Spiegel, S., & Kruglanski, A. W. (2002). Assimilative behavior identification can also be resource dependent: The unimodel perspective on personal-attribution phases. *Journal of Personality and Social Psychology, 83,* 542-555. doi:10.1037/0022-3514.83.3.542

Costa Jr., P., & McCrae, R. R. (1992). *NEO PI-R Professional Manual*.

Odessa, FL: Psychological Assessment Resources, Inc.
Denissen, J. J. A., & Penke, L. (2008). Motivational individual reaction norms underlying the Five-Factor model of personality: First steps towards a theory-based conceptual framework. *Journal of Research in Personality, 42,* 1285-1302. doi:10.1016/j.jrp.2008.04.002
Dirks, M. A., Treat, T. A., & Weersing, V. R. (2007). The situation specificity of youth responses to peer provocation. *Journal of Clinical Child & Adolescent Psychology, 36,* 621-628. doi:10.1080/15374410701662758
Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80,* 1011-1027. doi:10.1037/0022-3514.80.6.1011
Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2008). Integrating dispositions, signatures, and the interpersonal domain. *Journal of Personality and Social Psychology, 94,* 531-545. doi:10.1037/0022-3514.94.3.531
Fox, C. R., & Levav, J. (2004). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General, 133,* 626-642. doi:10.1037/0096-3445.133.4.626
Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford: Oxford University Press.
Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117,* 21-38. doi:10.1037/0033-2909.117.1.21
Gresham, F. M., Cook, C. R., Collins, T., Rasethwane, K., Dart, E., Truelson, E. et al. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the social skills rating system-teacher form. *School Psychology Review, 39,* 364-379.
Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system manual.* Circle Pines: American Guidance Service.
Hartley, A. G., Zakriski, A. L., Wright, J. C. (2011). Probing the depths of informant discrepancies: Contextual influences on divergence and convergence. *Journal of Clinical Child & Adolescent Psychology, 40,* 1-13. doi:10.1080/15374416.2011.533404
Helson, R., Jones, C., & Kwan, V. S. Y. (2002). Personality change over 40 years of adulthood: Hierarchical linear modeling analyses of two longitudinal samples. *Journal of Personality and Social Psychology, 83,* 752-766. doi:10.1037/0022-3514.83.3.752
Henry, D. B. (2006). Associations between peer nominations, teacher ratings, self-reports, and observations of malicious and disruptive behavior. *Assessment, 13,* 241-252. doi:10.1177/1073191106287668
Hoffenaar, P. J., & Hoeksma, J. B. (2002). The structure of oppositionality: Response dispositions and situational aspects. *Journal of Psychology and Psychiatry and Allied Health Disciplines, 43,* 375-385.
Hunsinger, M., Isbell, L. M., & Clore, G. L. (2011). Sometimes happy people focus on the trees and sad people focus on the forest: Context-dependent effects of mood in impression formation. *Personality and Social Psychology Bulletin*.
John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—Versions 4a and 54*. Berkeley, CA: University of California.
Kammrath, L. K., Mendoza-Denton, R., & Mischel, W. (2005). Incorporating if … then … personality signatures in person perception: Beyond the person-situation dichotomy. *Journal of Personality and Social Psychology, 88,* 605-618. doi:10.1037/0022-3514.88.4.605
Mischel, W. (2009). From personality and assessment (1968) to personality science. *Journal of Research in Personality, 43,* 282-290. doi:10.1016/j.jrp.2008.12.037
Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89,* 730-755. doi:10.1037/0033-295X.89.6.730
Reeder, G. D., Monroe, A. E., & Pryor, J. B. (2008). Impressions of Milgram's obedient teachers: Situational cues inform inferences about motives and traits. *Journal of Personality and Social Psychology, 95,* 1-17. doi:10.1037/0022-3514.95.1.1
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10). New York: Academic Press.
Schaller, M. (1992). In-group favoritism and statistical reasoning in social inference: Implications for formation and maintenance of group stereotypes. *Journal of Personality and Social Psychology, 63,* 61-74. doi:10.1037/0022-3514.63.1.61
Schwarz, N., & Oyserman, D. (2011). Asking questions about behavior: Self reports in evaluation research. In Melvin, M., Donaldson, S., & Campbell, B. (Eds.), *Social Psychology and Evaluation*. New York: Guildford Press.
Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: Distributed social cognition. *Psychological Review, 116,* 343-364. doi:10.1037/a0015072
Smith, R. E., Shoda, Y., Cumming, S. P., & Smoll, F. L. (2009). Behavioral signatures at the ballpark: Intraindividual consistency of adults' situation-behavior patterns and their interpersonal consequences. *Journal of Research in Personality, 43,* 187-195. doi:10.1016/j.jrp.2008.12.006
Tellegen, A. (1991). Personality traits: Issues of definition, evidence and assessment. In W. Grove, & D. Cicchetti (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (pp. 10-35). Minneapolis: University of Minnesota Press.
Terracciano, A., McCrae, R. R., & Costa Jr., P. (2009). Intra-individual change in personality stability and age. *Journal of Research in Personality, 44,* 31-37. doi:10.1016/j.jrp.2009.09.006
Trope, Y., & Gaunt, R. (2000). Processing alternative explanations of behavior: Correction or integration? *Journal of Personality and Social Psychology, 79,* 344-354. doi:10.1037/0022-3514.79.3.344
Vansteelandt, K., & Van Mechlen, I. (1998). Individual differences in situation-behavior profiles: A triple-typology model. *Journal of Personality and Social Psychology, 75,* 751-765. doi:10.1037/0022-3514.75.3.751
Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38,* 319-350. doi:10.1016/j.jrp.2004.03.001
Wood, D., & Roberts, B. W. (2006). Cross-sectional and longitudinal tests of the personality and role identity structural model (PRISM). *Journal of Personality, 74,* 779-810. doi:10.1111/j.1467-6494.2006.00392.x
Wright, J. C., Lindgren, K. P., & Zakriski, A. L. (2001). Syndromal versus contextualized personality assessment: Differentiating environmental and dispositional determinants of boys' aggression. *Journal of Personality and Social Psychology, 81,* 1176-1189. doi:10.1037/0022-3514.81.6.1176
Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53,* 1159-1177. doi:10.1037/0022-3514.53.6.1159