J Psychopathol Behav Assess (2011) 33:215–227
DOI 10.1007/s10862-011-9233-x

Reassessing the Assessment of Change in At-Risk Youth: Conflict and Coherence in Overall Versus Contextual Assessments of Behavior

Jack C. Wright · Audrey L. Zakriski · Anselma G. Hartley · Harry W. Parad

Published online: 15 May 2011

Abstract This research examined how a contextual approach to personality assessment can reveal change processes that are obscured by measures of overall behavior frequencies. Using field observations of 336 children from three summers at a program for at-risk youth, we illustrate how children's social experiences change over time, how their reactions to these experiences change, and how both processes contribute to changes in the overall frequencies of their prosocial, aggressive, and withdrawn behavior. Children showing opposite patterns of change in their environments and their reactions to them were nevertheless similar in their overall amount of change. The results clarify how changes in reactions and social experiences can be disentangled and reintegrated in order to deepen our understanding of personality change processes. Implications for change assessments that rely on overall behavior summaries are highlighted for program, individual, and intra-individual levels of analysis.

Keywords Behavior change · Personality processes · Individual differences · Assessment of children and adolescents · Aggression · Social withdrawal · Prosocial behavior

We are deeply grateful to the children, parents, staff, and administrators of Wediko Children's Services, whose cooperation made it possible to collect the data reported here. We would also like to thank the research coordinators and assistants for their dedication to the project. The first and second authors made equal contributions to this work; the order of their authorship was randomly determined. This research was partially 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.

J. C. Wright (✉)
Walter S. Hunter Laboratory of Psychology, Brown University,
89 Waterman Street,
Providence, RI 02912, USA
e-mail: Jack.C.Wright@gmail.com

A. L. Zakriski
Department of Psychology, Connecticut College,
270 Mohegan Avenue,
New London, CT 06320, USA
e-mail: alzak@conncoll.edu

A. G. Hartley
Walter S. Hunter Laboratory of Psychology, Brown University,
89 Waterman Street,
Providence, RI 02912, USA
e-mail: Anselma_Hartley@brown.edu

H. W. Parad
Wediko Children's Services,
72-74 East Dedham St.,
Boston, MA 02118, USA
e-mail: hparad@wediko.org

The study of behavior change in children has been hampered by a gap between theory and practice. On the one hand, researchers advocate for attention to social interactional and situational factors that reveal individual differences in behavior problems and competencies (Dirks et al. 2007a, b), influence behavior across settings (De Los Reyes and Kazdin 2005; Drabick et al. 2008), and mediate change and its generalization (Kazdin 2006). On the other hand, assessment practices rarely include process-oriented, context-sensitive methods, and instead emphasize overall behavior summaries. Mash and Hunsley (2005) note that despite the trend toward evidence-based treatment, assessment methods are seldom designed with specific treatment sensitivity or surrounding context in mind. Others note how

an emphasis on symptom change may conflict with impressions of improvement (Connor et al. 2002) and contribute to the finding of treatment equivalence when different mechanisms and individual responses are involved (Barlow and Nock 2009). Likewise, personality psychologists have expressed dissatisfaction with the reliance on nomothetic summaries that neglect situational variation and oversimplify personality processes (Mischel 2009; Van Mechelen 2009). In light of these concerns, the present research examines the conceptual foundations and consequences of acontextual assessment, and evaluates an alternative contextual approach to the study of change.

The predominance of summary approaches is rooted partly in personality psychology traditions, which have long emphasized nomothetic views of traits as enduring and cross-situationally consistent behavior tendencies (see Cervone et al. 2001). Although some personality researchers advocate for detailed process-oriented accounts of how people react to and influence their environments (Mischel 2009), widely used tools for assessing adult personality provide little explicit treatment of social situations or how the rater should take them into account (McCrae and Costa 1999). Instead, dispositions are operationalized as overall act frequencies that are not formally concerned with the immediate context surrounding an 'act' (Block 1989), and trait measures continue to focus on people's average tendencies to exhibit a given class of behavior (Cervone et al. 2001; Hershberger et al. 1995). Variability over situations is treated as measurement error, implying that people's stable and enduring characteristics are best revealed by aggregation across situations and occasions (Barkley 1988). "Complex patterns of behavior" are recognized, but it is simultaneously argued that these patterns "need to be examined, interpreted, and aggregated across numerous situations, places and times, to arrive at a reliable and valid index of a personality trait" (Roberts and Caspi 2001, p. 105).

This aggregationist view is exemplified in widely used child assessment methodologies. Retrospective rating methods (Achenbach and Rescorla 2001; Reynolds and Kamphaus 2002) often focus on how often a child displays various behaviors (e.g., "argues a lot", "teases a lot"), leaving the informant to judge the role of surrounding social events. Some items refer to events (e.g., "argues when denied own way", "feels hurt when criticized"), but the instruments do not examine how often these events occur or how they may influence behavior. Likewise, direct observational methods are also often used to assess overall behavior frequencies (see Volpe et al. 2005; Weisz et al. 2005) rather than to provide detailed analyses of proximal situational determinants (Nock and Kurtz 2005). The shared emphasis on aggregation is evident in the fact that several direct observation methods use behavior codes that parallel those in retrospective methods, and generate overall behavior counts to then validate the retrospective reports (Achenbach and Rescorla 2001; Ladd and Profilet 1996). The assessment of behavior change proceeds along similar lines. Summary behavior ratings or direct observations at two or more points are compared to assess the amount of change (August et al. 2001; Connor et al. 2002; NICHD 2003) without explicit consideration of behavioral variability over situations, and often with the assumption that the change being assessed is in the child rather than in the social environment.

Advocates of functional behavior assessment (Haynes et al. 2009), transactional approaches to development (Lewis 2001), and cognitive social learning approaches to personality (Cervone et al. 2001) have raised questions about what may be lost when researchers rely too heavily on broad context-free summaries. Related work on so-called contextual models of personality incorporate context into the assessment process by examining *if…then* relationships between social events and people's behavioral responses to them (Mischel and Shoda 1995; Van Mechelen 2009; Wright and Mischel 1987). For example, Vansteelandt and Van Mechelen's (1998) situation-response analysis of hostility revealed distinctive patterns of *if…then* links adults showed between antecedent events (e.g., *if* frustrated, ignored) and their hostile behavior (e.g., *then* attack, curse). Contextual models thus focus on the conditional probabilities of how a person responds to relevant conditions, or p(Behavior | Event), and what these conditional probabilities reveal about psychosocial functioning beyond the overall trait scores (Fournier et al. 2008; Smith et al. 2009). Research using this approach has shown how analyses of situation-behavior profiles can clarify individual differences in several adult and child behaviors (Hoffenaar and Hoeksema 2002; Van Mechelen and Kiers 1999).

One implication of this work is that overall measures are insensitive to distinct processes that contribute to behavior (Dirks et al. 2007a; Scotti et al. 1996). People who display similar overall frequencies or symptom ratings, p(Behavior), can differ in the conditional probabilities of their responses to events, p(Behavior | Event). For example, boys who are equally high in overall externalizing behavior show distinct patterns of responses to aversive and nonaversive events in interactions with peers and adults (Wright and Zakriski 2001). Overall measures can conflate variation in how often children encounter events, p(Event), with variation in how they react to them. Boys who are often provoked by peers, but unlikely to respond aggressively when this occurs, may appear similar to boys who are seldom provoked, but likely to respond aggressively when this occurs (Wright et al. 2001). Even when individuals differ in their overall frequencies, they may show the same responses to events, and instead differ only in how often they encounter them (Zakriski et al. 2005).

Contextualized personality research often focuses on behavior patterns individuals or groups show in response to a canonical set of stimuli (Fournier et al. 2008; Noftle and Fleeson 2010; Smith et al. 2009). This approach intersects with but is distinct from other work that incorporates context. Functional behavior assessment typically involves case studies in which the stimuli and/or outcomes are idiosyncratically defined for the individual (Haynes et al. 2009). Some social competence research examines taxonomies of situations, but assesses "inappropriate behavior" in each situation rather than children's specific behavioral reactions (Dodge et al. 1985; Matthys et al. 2001). Multisituational assessment methods (Lutz et al. 2002; McDermott 1993) examine specific responses to situations (e.g., aggression when answering teachers' questions, handling peer conflict), but these are then aggregated to assess cross-situationally pervasive behavior problems (e.g., "oppositional defiant"). Perhaps the most common approach is to mix specific event-reaction pairings ("responds to teasing or name calling by ignoring"), and uncontextualized items ("shows sympathy for others"), which are aggregated into omnibus scales (e.g., Gresham and Elliot 1990; Measelle et al. 2005; Walker and McConnell 1995). This blending of global and contextualized items makes it difficult to disentangle the intraindividual and environmental processes that contribute to ostensibly "dispositional" constructs.

Past research illustrates advantages of contextual assessment for understanding personality processes, but has given little attention to the question of how people change. One critical task is to clarify conditions under which overall and context-specific measures yield similar conclusions about change; when this occurs parsimonious summary measures could suffice. Another is to understand whether some context-specific changes are too complex to afford simple summaries and whether some summary measures conflate conceptually distinct changes in children's responses to events and changes in their social environments. A third is to probe the social interactional processes that contribute to "change-sensitive" measures that are context-specific (e.g., "responds appropriately when hit/pushed") versus uncontextualized (e.g., "overall classroom behavior") (see Gresham et al. 2010).

To explore these issues, we examine changes in children's aggressive, withdrawn, and prosocial responses to multiple social events (peer approach, tease, and bully; adult praise, instruct, and warn) during summer treatment. We focus on these behaviors because they include common referral behaviors in treatment and school settings (e.g., Connor et al. 2002; Dirks et al. 2007a), and are found in a wide range of child behavior inventories (e.g., Achenbach and Rescorla 2001; Ladd and Profilet 1996; Reynolds and Kamphaus 2002). We focus on these events because they are common in children's interactions, present a range of interpersonal demands, and elicit meaningful inter- and intra-individual differences in behavior (Dirks et al. 2007a, b; Shoda et al. 1994; Zakriski et al. 2005). We use extensive field observations of behavior because this methodology reduces biases associated with retrospective ratings and is sensitive to the context-dependent organization of behavior (Funder 2009; Furr 2009; Kagan 2001). We extend past research by simultaneously examining changes in children's overall behavior rates, changes in their reactions to events, and changes in how often they encountered those events.

The lack of evidence on simultaneous changes in overall behavior rates, events rates, and context-specific reactions dictated caution in our predictions. First, we hypothesized that overall and context-specific measures would yield diverging conclusions about change. Reactions assess change in the conditional probabilities of responses given that certain events are encountered, p(Behavior | Event), but deliberately avoid assessing changes in the frequency of those events, p(Event). Overall measures assess behavior frequencies over a period of observation, p(Behavior), and so are influenced by both. Depending on changes in how often children encounter events that elicit aggression, they could show a decrease in their overall rate of aggression, yet also show an increase in the conditional probability of their aggressive reactions.

A second, related hypothesis is that children will show diverse changes in their reactions to different events, with ostensible improvement in response to some and worsening in response to others. Such divergences are especially interesting because they may cause aggregated measures to miss important but conflicting context-specific changes. Past work led us to expect that increases in prosocial behavior and decreases in problem behavior (i.e., aggression, withdrawal) would be greatest in response to nonaversive events (e.g., peer/adult talk) that are unlikely to lead to coercive escalation (Granic and Patterson 2006). We expected reactions to conflict situations to be more resistant to change, or even show adverse change, as these situations make greater demands on children's self-regulatory skills (Dirks et al. 2007b; Wright and Mischel 1987) and create opportunities for deviant peer influence (Dodge et al. 2006).

Third, we examine processes that contribute to convergences and divergences between context-specific and overall measures of change. In contrast to studies that blend contextualized and uncontextualized items into omnibus scales, we probe how events, reactions, and overall measures are distinct yet coherently inter-related. We expected that aggregation of reactions over increasingly broad samples of events would not necessarily converge with overall measures because the latter are also influenced by how often events are encountered. We then test an alternative method that combines event-specific reactions and the frequencies of the events

themselves. We expected this "integration" method to converge better with what is assessed by overall measures. Such evidence would reinforce the view that summary measures, rather than being unambiguous indicators of the person, are person-situation mosaics requiring careful deconstruction.

Fourth, we test whether changes in children's reactions and in the events they encounter predict individual differences in their overall change. We expected event change and reaction change to be modestly related, but that each would make unique contributions to predictions of overall change. We expected that even children with opposite functional change patterns would be hard to distinguish based on their overall change. Thus, children who show decreases in aggressive reactions to aversive events, but who encounter increases in those events, are functional opposites of children who show increases in their aggressive reactions but encounter decreases in those events. Yet, if overall assessments of behavior change are indifferent to such change processes, children showing these opposite patterns should be comparable in overall change. Finally, using idiographic profile analysis methods (Mischel and Shoda 1995; Smith et al. 2009), we further illustrate how changes in individual children's reaction profiles and event profiles can be used to disambiguate children who show comparable changes in overall behavior.

Method

The study was part of a multiyear project at a residential program for at-risk youth. The program serves 120–140 children each summer, referred for behavioral, academic, and social problems. Children are primarily from urban public schools in New England, but also from other suburban and rural schools in the region. Other research in this setting reports that these percentages of children met clinical cutoffs (T-score≥70): 53% and 27% for parent presummer CBCL aggression and withdrawal, respectively (Hartley et al. 2011); 53% and 36% for counselor in-summer TRF externalizing and internalizing, respectively (Wright et al. 1999). Children live in the setting for 45 days each summer in groups of 8–10 same-sex, same-aged peers. Each day, their schedule includes classroom instruction, structured activities (e.g., academics, art, swimming), and group therapy (see Hartley et al. 2011; Zakriski et al. 2005). Research permission was obtained from parents/guardians during the interview process.

Participants A total of 336 children drawn from three summers met inclusion requirements (see below); 51% were White, 37% African American, 9% Hispanic, 1% Asian, and 2% other; children were predominantly lower- and middle-SES. Analyses used age groups (<12; >= 12 years) (see Achenbach and Rescorla 2001), with *Ns* of 135 (younger boys), 92 (older boys), 66 (younger girls), and 43 (older girls). Of the 457 children attending during the summers studied, those arriving late or with fewer than 75 observations were excluded; these were primarily adolescents in prevocational groups whose schedules did not permit frequent observations. Teachers and activity counselors ran classes and activities 4 h/day; residential counselors were with their group most of the day. Non-supervisory staff (N=246) provided observational data described next.

Behavior Coding Materials and Procedure

We used a coding system similar to one described elsewhere (Mischel and Shoda 1995; Zakriski et al. 2005). Ten codes assessed overall behavior rates and reaction rates (see preliminary analyses): "hit, pushed, physically aggressed," "teased, provoked or ridiculed," "bossed, bullied, or threatened," "argued or disapproved," "withdrew, isolated self," "whined or cried," "talked in age-appropriate way," "attended or listened to other(s)," "showed positive emotion," and one too infrequent to use, "self-stimulation/self-abuse." Seven codes assessed whether specific social events occurred: "adult praised the child verbally;" "adult gave the child a warning;" "adult instructed the child to do something;" "adult gave the child a time out" (a form of discipline); "peer talked in age-appropriate way;" "peer teased, provoked, or ridiculed;" "peer bossed, bullied, or threatened." Detailed definitions were provided during training.

On each optical form, the coder rated how often a child displayed each behavior ("On the whole, how did the child behave during this observation period?") using a 0–3 scale ("not at all," "somewhat," "moderately", "a lot"). The coder next reported whether the target encountered each social event, identified the peer or adult involved, and then rated the target's reaction to that event using the same behavior codes and 0–3 scale just noted. If an event occurred more than once, coders reported only the most recent one. Multiple reactions were allowed to a given event.

Adults assessed children at the end of hour-long activities. Each adult completed forms for 2–6 children/period, and each child was observed in 3–5 periods/day (except Sundays), yielding 36 coding days/summer. Children were rated by many adults, including counselors who interacted with them throughout the day, and teachers who worked with them primarily in their classroom.

Preliminary Analyses

As in Zakriski et al. (2005), scales were formed from individual codes. Nine behaviors were combined into 3

scales: *aggression* (argue, tease, boss, and hit/attack), *withdrawal* (withdraw, whine), and *prosocial* (talk, attend, positive emotion). We label these AGG, WDR, and PRO, respectively. For events, adult warn and discipline were aggregated, as were peer tease and peer boss. This yielded 5 categories, labeled as follows: *adult warn/discipline* (AWND), *peer tease/boss* (PTEB), *adult praise* (APRA), *adult instruct* (AINS), and *peer talk* (PTLK).

The "overall behavior rate" was the average overall behavior rating a child received on each day. The "event rate" was the number of times a child encountered each event on each day divided by the total number of events for that child that day. Using reaction ratings recoded to binary form (0 = no response, 1 = non-zero rating, as previously noted), the "reaction rate" was the number of times a response was displayed in reaction to a specific event divided by the number of times that event occurred (e.g., # instances of withdraw to tease/boss divided by # instances of being teased/bossed). This yielded 3 (reaction type) x 5 (event type) or 15 event-reaction pairs.

Data were pooled over summers, yielding 42,471 hourly records, 76,433 contexts, and 140,150 reactions. Mean overall behaviors (0–3 scale) were: aggression (.41), withdrawal (.58), and prosocial (1.75). Mean event rates (0–1 range) were: adult instruct (0.32), adult praise (0.25), adult warn/discipline (0.20), peer talk (0.14), and peer tease/boss (0.09). Mean reactions to events (0–1 range) were: adult warn/discipline (AGG=.39, WDR=.18, PRO=.42), peer tease/boss (.63, .18, .19), adult instruct (.12, .09, .79), peer talk (.07, .04, .89), and adult praise (.01, .03, .95). Aggressive and withdrawn reactions to positive events were infrequent (see Mischel and Shoda 1995) and prosocial reactions to these events were frequent (Wright and Zakriski 2001).

Reliability analyses were similar to those reported elsewhere (Zakriski et al. 2005). Analyses of individual coders would have been uninformative because each coder provided relatively few observations and did not code all children. We therefore examined split-half reliability using odd and even days. For overall rates, Spearman-Brown reliabilities were: aggression (.93), withdrawal (.91), and prosocial (.92). For event rates the results were: adult warn/discipline (.80), adult praise (.75), peer tease/boss (.73), adult instruct (.69), and peer talk (.65). Based on Zakriski et al. (2005), we expected reliabilities to be highest for reactions to aversive events. These were: to adult instruct (AGG=.82, WDR=.64, PRO=.78), to peer tease/boss (.83, .84, .56), and to adult warn/discipline (.72, .71, .76). Ceiling/floor effects noted previously for reactions to positive events attenuated their reliabilities: peer talk (AGG=.50, WDR=.55, PRO=.59), and adult praise (.44, .42, .35). We report all reactions because they are needed to interpret overall behaviors.

All mean *r*s reported here are based on Fisher's *r*-to-*z* transform. Significance in repeated-measures analyses was based on Greenhouse-Geisser adjustments (Maxwell and Delaney 1990).

Results

We first performed "program level" analyses of mean behaviors, reaction rates, and event rates on individual days because this provided the observations per day needed to measure daily change reliably and assess whether linear models were appropriate. Guided by these results, we then performed analyses that examined individual differences in change. To measure each child reliably, these analyses used time intervals that aggregated over multiple days within child.

Mean Change over Days

Using means for each day (N=36) for each summer, we examined children's overall behavior rates. We used multilevel modeling (Bryk and Raudenbush 1992) to examine whether intercepts and slopes varied over years. For each behavior, AIC was smaller (better) for the simpler fixed-effects model compared to the random intercept model, χ^2s(1)>5.40, ps<.02, and modestly smaller compared to the model with random slopes and intercepts, χ^2s(3)>6.70, ps<.09. Because no advantage was found for models with random effects, we computed the mean daily behavior over children and years. Figure 1 shows the results in standardized form (M=0, SD=1), including linear and non-linear fits (local least-squares, Chambers and Hastie 1992). Aggression and withdrawal showed significant linear decreases; prosocial behavior showed a linear increase. The non-linear fit was better only for prosocial behavior. The few non-linearities we found for these or other measures were due to perturbations very early or late in the session; subsequent analyses examined linear change after removing the first 2 and last 2 days, leaving 32 days.

Linear change over days can be summarized as the correlation (r) between mean behavior rate and day, and the slope (b) of unstandardized behaviors regressed on day. As shown in Table 1, we found significant decreases in overall aggressive and withdrawn behavior, and significant increases in prosocial behavior. Thus, if overall behavior rates were the criterion, children "improved", as they showed decreases in problem behavior and increases in prosocial behavior.

As hypothesized, changes for event-specific reactions were variable and at times conflicted with those for overall behavior rates (see Table 1). Aggressive reactions increased

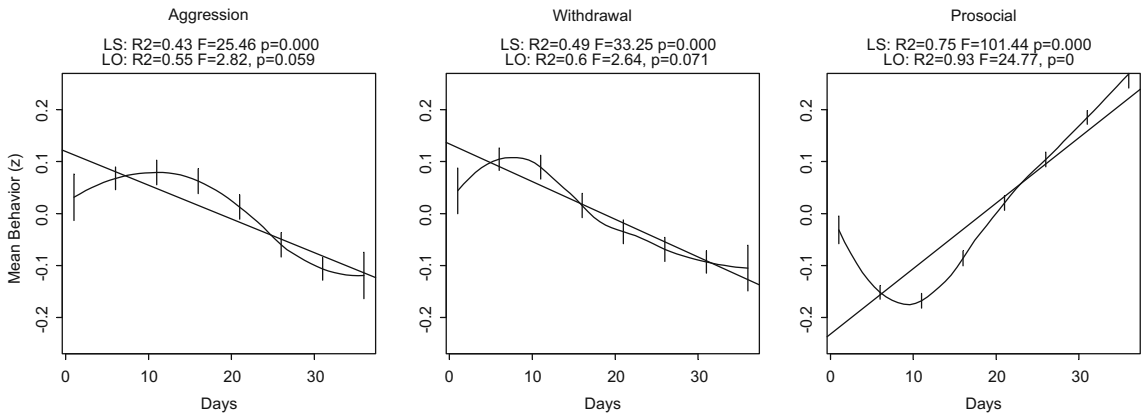

Fig. 1 Change over days for mean overall behavior rates (z-scores). Least-squares regression lines are shown, with model summaries atop ("LS"). Non-linear, local polynomial least-squares solutions, with simultaneous 68% confidence intervals, are superimposed. Comparisons of non-linear and linear models are provided above each panel ("LO")

in response to peer tease/boss and to peer talk, but showed little change elsewhere. Prosocial reactions to adult praise and instruct increased, but decreased for peer tease/boss. Contrary to predictions, changes in withdrawn reactions varied little over events, with reliable decreases for 4 of the 5 events.

To examine how change results were affected when reactions were aggregated over multiple events, we computed change for each level of aggregation that was possible using the five events. For the lowest level (1), we computed the mean of the 5 single-event *r*s in Table 1. For the next (2), we created the 10 possible combinations of 2 events (e.g., aggressive reactions to peer tease/boss and to adult warn/discipline). For each combination, we computed the mean reactions to those events on each day and then computed the *r* between the mean reactions and days. We accumulated the 10 *r*s, then computed the mean *r* at this level. We used the same method for all sets of 3 (6 combinations), 4 (3 combinations), and 5 (1 combination, all) events.

For aggressive and prosocial reactions, change results based on aggregated reactions did not converge with those based on overall rates. As Fig. 2 (left) shows, *r*s for aggressive reactions were positive, indicating increases over days. These results reflect the contribution of increases in aggressive reactions to peer events. Even at maximum aggregation the *r* differed from the overall rate result in Table 1 (−.64; see "OR" in Fig. 2). Mean *r*s for prosocial reactions were near 0 regardless of aggregation, reflecting the variability in individual prosocial reactions and again contrasting with the overall result. Mean *r*s for withdrawn reactions essentially summarize the consistent decreases already reported for individual withdrawal reactions and also resembled the overall result. We return to the second panel of Fig. 2 after explaining the analysis.

Before examining how reactions and events combined, we examined how events themselves changed. Peer tease/boss and adult warn/discipline decreased (*rs*=−.61, −.45, *ps*<.01); adult instruct and peer talk increased (.43, .38, *ps*<.05).

Table 1 Linear change over days in children's overall behavior rates and event-specific reactions

			Event-specific reactions				
Behavior/Reaction	Coeff.	Overall rate	APRA	PTLK	AINS	AWND	PTEB
Aggression	*r*	-.64***	.12	.47**	-.19	.17	.65***
	b[a]	-.08	.01	.03	-.01	.02	.12
Withdrawal	*r*	-.77***	-.59***	-.30	-.53**	-.42*	-.48**
	b	-.12	-.02	-.02	-.03	-.03	-.06
Prosocial	*r*	.90***	.51**	-.19	.45**	.18	-.51**
	b	.22	.02	-.02	.04	.02	-.06

APRA Adult praise, *PTLK* Peer talk, *AINS* Adult instruct, *AWND* Adult warn/discipline, *PTEB* Peer tease/boss

[a] Redundant significance flags for slope (*b)* are not shown.

*p<.05; **p<.01; ***p<.001

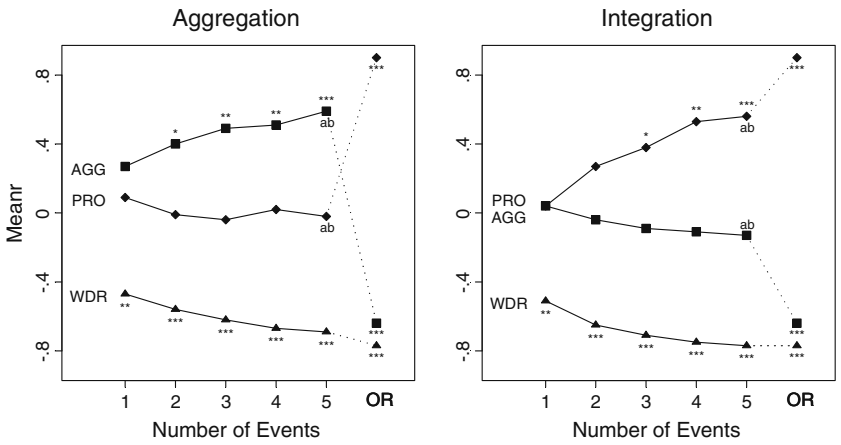

Fig. 2 Results for aggregation and integration measures of change. Aggregation method (left) shows change over days (indexed as *r*) in mean conditional probability of reactions to events, as a function of the number of events assessed. Integration method (right) shows change in marginal probabilities computed from the probability of each reaction and the probability of encountering each event. *AGG* aggression; *WDR* withdrawal; *PRO* prosocial behavior. "OR" indicates the result based on the overall rate measure; its position on the abscissa is arbitrary. Asterisks provide significance tests against $\rho=0$; * $p<.05$; ** $p<.01$; *** $p<.001$. [a] $p<.01$ for comparison with corresponding *r* for OR; [b] $p<.01$ for comparison with corresponding *r*s for aggregation versus integration methods

Adult praise showed little change (.23). Events and reactions combined in various ways. For example, the probability of aggression to peer tease/boss was high ($M=.66$) and increased ($r=.65$, $b=.11$, $p<.001$), but the probability of this event was low ($M=.10$) and decreased ($r=-.61$, $b=-.03$, $p<.001$). The joint probability of observing this event and this reaction, p(AGG | PTEB) * p(PTEB), was low ($M=.06$) and showed little change ($r=-.23$, $b=-.01$, $p>.20$). For prosocial reactions to adult instruction, both p(PRO | AINS) and p(AINS) were relatively high (Ms=.79, .32), both increased (rs=.45, .43, bs=.03, .02, ps<.02), and the joint probability increased ($M=.25$, $r=.55$, $b=.03$, $p<.01$).

We calculated the integration measure by incorporating multiple events. For example, for aggression to peer tease/boss and to adult instruct, the "marginal" is: p(AGG | PTEB) * p(PTEB) + p(AGG | AINS) * p(AINS). As Fig. 2 (right) shows, for aggression, *r*s for integration were near 0 regardless of the number of events. For prosocial behavior, *r*s increased with the number of events. For both behaviors, the *r* using integration more closely resembled the *r* for overall rates than the aggregation method did, highlighting the influence of both event change and reaction change on overall measures. For withdrawal, *r*s for integration and aggregation were similar.

Within-Subject Analyses

Analyses of program-level change do not reveal individual differences in change or their moderators (e.g., age). We therefore performed analyses with children as units of analysis. To ensure reliability, we formed two time periods (16 days each) and computed means for each child per period. Multilevel modeling did not reveal improvement when intercepts or slopes (over years) were treated as random effects, all χ^2s(3)<4, ps>.2. Moreover, the data contained no missing cases and time points were invariant. We therefore performed fixed-effects ANOVAs, with age and gender as between-subjects factors and time (and event where needed) as within-subjects. We report the few age and gender effects we found after the primary analyses.

Results for overall rates paralleled program-level findings already presented. There was a time x behavior interaction, $F(2, 664)=69.20$, $p<.001$. Aggression and withdrawal decreased, Fs(1, 332)=21.02, 52.26, ps<.001, and prosocial behavior increased, $F(1, 332)=90.90$, $p<.001$.

The aggregated reactions also showed a time x behavior interaction, $F(2, 664)=10.27$, $p<.001$. In contrast to the results just noted based on overall rates, aggressive reactions increased, $F(1, 332)=6.91$, $p<.01$, and prosocial reactions did not change, $F<1$. Withdrawn reactions decreased, $F(1, 332)=23.50$, $p<.001$, resembling the overall rate result.

An analysis of event rates revealed a time x event interaction, $F(4, 1328)=8.75$, $p<.001$. Adult praise ($p<.02$) and peer talk ($p<.003$) increased; adult warn/discipline ($p<.01$) and peer tease/boss ($p<.001$) decreased. These results converge with the program level analyses.

Finally, we integrated reaction and event rates as before. A time x behavior interaction was found, $F(2, 664)=19.10$, $p<.001$. Aggression and withdrawal decreased, Fs(1, 332)=3.92, 32.09, ps<.05, .001, and prosocial behavior increased, $F(1, 332)=21.45$, $p<.001$. As expected, overall change results resembled the integration results more closely than the aggregation results, again showing the

Table 2 Correlational and multiple regression analyses of change in overall behavior, event rates, and reaction rates, by behavioral domain

	Correlation (*r*)			Multiple regressions			
Domain	E-O	R-O	E-R	E	R	*F*	R^2
Aggression	.40**	.43**	.16*	.34**	.38**	46.93	.30
Withdrawal	.33**	.40**	.20**	.26**	.35**	31.99	.22
Prosocial	-.44**	.37**	-.28**	-.36**	.28**	38.72	.26

O change in overall behavior rate; *E* change in the rate of aversive events; *R* change in the rate of reactions to aversive events. Degrees of freedom for *F* are 3, 332.

*$p<.01$; ** $p<.001$

influence of event change and reaction change on the overall rate measures.[1]

To clarify how changes in events and reactions predicted overall change, we performed multiple regressions using change scores (Time 2 – Time 1). Guided by stepwise regressions showing which variables predicted overall change,[2] we computed change in mean reactions to all "aversive" events (AWND, AINS, PTEB), and change in the rate of those events. As shown in Table 2, all *r*s with overall change were significant. For each domain, multiple regressions indicated that event and reaction change made unique contributions to predictions of overall change.

To test whether children with opposite patterns of event and reaction change were similar in overall change, we obtained predicted values at +/−1SD from regressions using event change, reaction change, and their interaction. Aggression change was lowest when both events and reactions decreased ($M=-.45$), highest when both increased (.46), and intermediate for offsetting pairs (.02, −.03). All comparisons were significant, $ts(332)>5.18$, $ps<.01$, except, as expected, the opposite pairs, $t<1$. For withdrawn and prosocial behavior, all comparisons were also significant, $ts(332)>3.50$, $p<.001$, except the expected opposite pairs, $t(332)<1.20$, $ps>.10$. Thus, children with distinct change processes did not differ in their overall change.

We next illustrate how reactions and events can be used idiographically to disambiguate pairs of children who were similar in their overall aggression change. Figure 3 (top row) presents two children who showed little overall aggression change (see label "O"), yet showed distinctive changes in their reactions to events. Child 57 (panel 1) became less aggressive in response to aversive events (A−, P−), but more aggressive in response to peer talk (P+). Child 116 (panel 3) showed the opposite pattern, becoming less aggressive in response to nonaversive events (A+, P+, Ai), but more aggressive in response to aversive ones. These children also differed in the event changes they experienced. Child 57 (panel 2) experienced a decrease in P+, but an increase in P−. Child 116 (panel 4) experienced, among other changes, an increase in A+, but a decrease in P+. Row 2 presents two children who showed increases in their overall aggression, but who differed in how their reactions and events changed. Child 74 became more aggressive to instruction and aversive events (Ai, A−, P−; panel 1), and had a very stable event profile featuring elevated adult discipline (A−, panel 2). Child 200 became more aggressive to all but one event (panel 3), even though he encountered aversive events less often (A−, P−, panel 4). Children who showed comparable *decreases* in overall aggression also showed variation in their reaction and event changes (not shown in Fig. 3). Thus, measures of change in overall rates obscured variability in children's reaction changes, event changes, and in the interplay between the two.

[1] Although age and gender were not our main focus, and the small number of girls dictated caution, we summarize the results for difference scores (T2 - T1). For aggressive reactions, there was a main effect for age, $F(1, 332)=4.09$, $p<.05$; younger children showed smaller increases than older ones (*M*s=.08, .19). A gender x event interaction was found, $F(4, 1328)=2.59$, $p<.04$. Girls' aggression to peer talk increased (.24), whereas boys' decreased (−.12); the reverse was found for reactions to adult instruct (−.05, .07). For withdrawal, there was an age effect (*M*s=−.16, –.32 for younger vs. older), $F(1, 332)=6.21$, $p<.02$, and gender (*M*s=−.34, –.14, for girls vs. boys), $F(1, 332)=10.73$, $p<.001$, and a 3-way interaction, $F(4, 1328)=4.14$, $p<.005$. Older girls showed steeper decreases in withdrawal to adult warn/discipline than younger girls (*M*s=−.52, –.15), but not to peer tease/boss (−.12, –.24). The reverse was found for boys. For prosocial, there was an interaction between gender and type of reaction, $F(4, 1328)=3.11$, $p<.02$. Girls showed steeper increases than boys in their prosocial reactions to adult praise (.36, .13) and to adult instruct (.23, .03), whereas boys showed larger increases in their prosocial reactions to peer talk (.15, –.04).

[2] We performed stepwise regressions with changes in the 5 event rates and 5 reaction rates as predictors, entering age and gender at step 1, and using a .01 entry threshold for other predictors. Age and gender accounted for little variance ($R^2<.03$); older children showed less change in withdrawn and prosocial behavior. Both event change and reaction change predicted overall change; event change entered first for aggressive and prosocial behavior. For aggression change the predictors to enter (in order) were adult warn/discipline ($\Delta R^2=.23$), aggression to adult instruct (.09), aggression to adult warn/discipline (.05), and peer tease/boss (.02); R^2 for the full model was .40. For prosocial the predictors were adult warn/discipline ($\Delta R^2=.16$), prosocial to adult instruct (.07), and adult instruct (.02), and full-model R^2 was .28. For withdrawal the predictors were: withdrawal to adult instruct ($\Delta R^2=.17$), adult warn/discipline (.11), and withdrawal to adult warn/discipline (.02), and the full-model R^2 was .32.

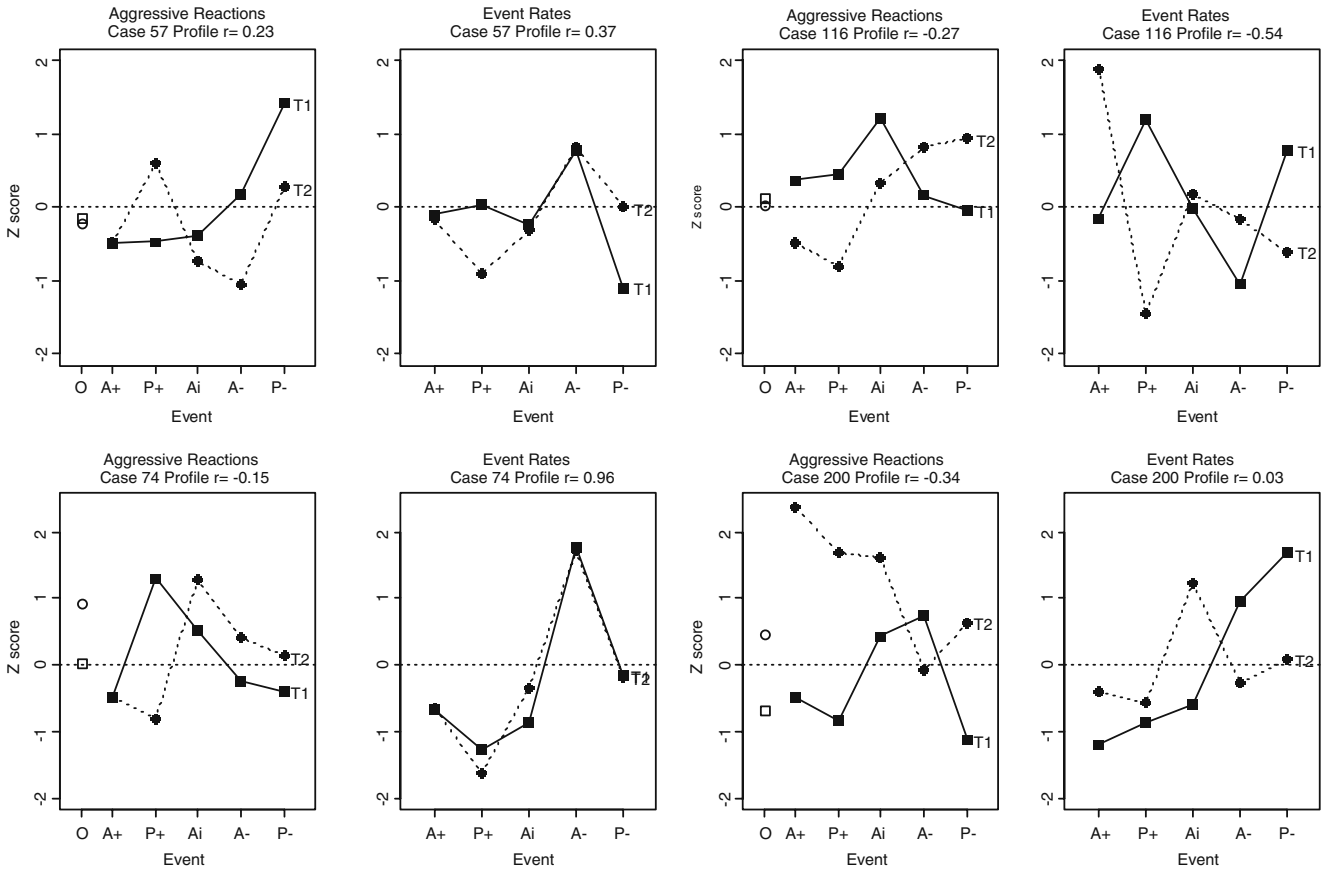

Fig. 3 Overall aggressive behavior, aggressive reactions to events, and rates of encountering events (z-scores), as a function of time for four children. T1 = Time 1; T2 = Time 2. O = overall frequency of behavior; A+ = adult praise; P+ = peer talk; Ai = adult instruct; A- = adult warn/discipline; P- = peer tease/boss. Profile *r* indicates correlation (Pearson's *r*) between profiles at time 1 versus 2, excluding the overall frequency measures ("O")

Finally, we examined intraindividual stability for all children using the type of profile *r*s shown in Fig. 3. Norming (see Method) ensured that these *r*s were not affected by the overall magnitudes (see Mischel and Shoda 1995). Mean profile stability was reliably above 0 despite variability over children. For aggressive, withdrawn, and prosocial reaction profiles, respectively, mean *r*s were .30 (SD=.50, range=−.94 to +.96), .17 (SD=.51, range=−.98 to +.99) and .27 (SD=.52, range=−.97 to +.99), *t*s(335)> 4.35, *p*s<.001. Stability for withdrawn reactions was lower than for the other reactions, *t*s(335)>2.03, *p*s<.05. The stability for events was relatively high (mean *r*=.48, SD=.46, range=−.91 to +.99), *t*(335)=13.71, *p*<.001, and exceeded each reaction result, *t*s(335)>4.01, *p*s<.001.

Discussion

This research answers questions about the nature of change, how to assess it, and what can be revealed by probing the social interactional processes that contribute to it. It responds to concerns about the gap between contextually rich theorizing about behavior and uncontextualized assessment practices (Dirks et al. 2007a), and builds on contextual models of personality to reveal individual differences in change processes. Five main findings emerged about summary and contextualized measures of change. First, we found clear change in overall rates of social behaviors: Aggression and withdrawal decreased, and prosocial behavior increased. If the "amount" of change were the criterion, the conclusion presumably would be that children "improved." Indeed, the overall change we found compares favorably with other studies (August et al. 2001; Connor et al. 2002) given that we used field observations rather than retrospective ratings.

Second, beneath this overall "improvement," we found diverse changes in event-specific reactions. Whereas the overall rate of prosocial behavior increased, children became more prosocial in response to adult instruction and praise, but less prosocial to peer provocation. Whereas overall aggression decreased, children became more aggressive in response to peers, especially to peer provocation. Changes in withdrawn reactions were more consistent across events and converged with the overall rate measure. In sum, it would be misleading

to reduce the pattern of children's reaction changes to summary statements about "improvement" or "worsening." Rather, children improved in some aspects of their social interactions, but worsened in others.

Third, we found little evidence that discrepancies between different measures of change were reduced by aggregating over reactions to multiple events. If anything, as more events were aggregated, discrepancies became clearer. For prosocial reactions, aggregation yielded no evidence of increases, which contrasted with the clear increases in prosocial behavior found for the overall rate measure. Children showed clear increases in their aggressive reactions even with maximum aggregation over events, despite equally clear decreases in the overall rate of aggression. Only for withdrawal did the aggregated reaction result and the overall rate result converge.

Fourth, although overall changes and reaction changes showed key discrepancies, coherence emerged when environment changes were taken into account. Children became more aggressive in response to peer provocation, but their rate of being provoked decreased, thereby attenuating change in the marginal rate of aggression. Children became more prosocial to adult instruction, and the rate of instruction also increased, thereby amplifying that change. When the probabilities of encountering events and the conditional probability of responses to them were integrated, we obtained results that corresponded better with changes in overall behavior than when the simple aggregation method was used. Stepwise regressions clarified that changes in both event rates and reaction rates predicted individual differences in children's overall behavior change. Indeed, changes in children's environments predicted overall change in aggressive and prosocial behavior better than did any change in their aggressive or prosocial reactions.

Fifth, the results highlight how different change processes can be missed by measures of overall change. Children who became more reactive to peer provocation and adult discipline, yet encountered decreasing rates of those events, showed overall change that was comparable to children who became less reactive, yet encountered increases in those events. Idiographic analyses clarified how individual children with comparable overall change can show differences in how their reactions and social experiences change over time. Individual children who showed little overall aggression change showed increased aggression in response to some events and decreased aggression in response to others. Children who showed comparable overall increases (or decreases) in their overall aggression also differed in locus and breadth of their reaction changes and in whether their social environments amplified or diminished these changes.

One might argue that our results stem from the particular, perhaps unrepresentative, events we sampled. Several points deserve note here. First, we sampled events children often encounter and that appear in various measures (Gresham and Elliot 1990; Matthys et al. 2001). Still, no one set of events can be equally representative of all children's social experiences. Children differ in the events they encounter and their appraisals of them; what is relevant for one child will not be for another. Taking each child's event frequencies into account partially addresses this, but only for the events that were sampled. Second, despite the limitations of the events and reactions we sampled, our results show how an analysis of their interconnections can be useful: Overall measures of change that are often presumed to assess the child instead reflect both changes in his reactions and in his social environment. Even when a context-specific reaction and an overall measure are correlated, they can reflect distinct psychosocial influences; whether these should be conflated into an overall score, as is done in scales that blend diverse correlated items, requires more than psychometric justification. Third, although our study does not rule out different results for other events, the changes we found still need to be understood. Contextually narrow adverse change (e.g., responses to peer conflict) could be critical to children's adjustment in school even if they also show other positive changes (e.g., to adult praise) that are more relevant at home. Future research should examine a wider range of events and reactions, probe the extent to which specific reaction changes generalize to home and school, and study how this is mediated by the events children experience in those settings.

Contextualized approaches raise certain issues that are less salient in summary methods. When events are rare, it is difficult to assess reactions reliably, but these reactions may be especially informative. Physical aggression is relatively infrequent (Tremblay 2000), yet children's reactions to it are particularly change-sensitive (Gresham et al. 2010). Likewise, rare reactions to common events (e.g., aggression to peer approach) may be revealing when they do occur (Kempes et al. 2010; Wright and Zakriski 2001). Our results do not imply that one should emphasize rare events or reactions, as contextualized assessments might, de-emphasize them, as "integration" (i.e., event-weighted) methods could, or merely leave it to the human judge to sort things out, as might occur when a "pure" uncontextualized item is used. Rather, our results imply that investigators should be aware that each of these strategies is based on different assumptions about what needs to be measured and that each may have distinct consequences for the changes that are detected or missed.

Our methodology also raises questions about reciprocal influences between children's "reactions" and the "events" they encounter, which in this study were behaviors of other people. This is especially complex when children are both the targets of treatment and the social stimuli to which other children respond. Future research will need to study the nested macro- and micro-settings (e.g., classroom, treatment group) to clarify the mechanisms by which interventions have their

effects (Zakriski et al. 2011). Although our field study necessarily leaves many causal questions unanswered, it reinforces calls to move beyond informants' summary ratings of traits and to be increasingly explicit about the rating process and the context-dependent organization of behavior (Cervone et al. 2001; Furr 2009; Hartley et al. 2011).

Our focus on overall behavior change stems from the emphasis in past research on such measures and does not imply that these measures are the gold standard for evaluating others. On the contrary, our results emphasize the need to evaluate the suitability of any measure for a given context and purpose (Mash and Hunsley 2005). Overall measures can be useful when the total behavior rate is of interest, regardless of its origins. A school's objective might be to reduce antisocial behavior, whether this is done by altering the environment, children's reactions, or both. Conditional responses to events would be useful when it is important to disambiguate events, factor out changes in how often events are encountered, and make inferences about the child as the locus of change. Whether aggregating reactions over multiple events is informative will depend on the variability of children's reactions, as our results for prosocial versus withdrawn behavior show. A simultaneous analysis of reaction change and environment change, as illustrated here, may be especially important when there is reason to believe that the processes interact.

One might argue that changes in a child's environment are merely indirect effects of changes in her behavior, and thus there is no need to separate the two when assessing overall change. This oversimplifies the evidence. Change in aversive events was only modestly correlated with changes in children's reactions (rs=−.28 to .20); event change and reaction change made unique contributions to changes in overall behavior. Overall measures missed important effects when changes in children's reactions were offset by changes in their environments. Children who show increases in their aggressive reactions to provocation, but who are provoked less often are distinct from children who show *decreases* in such reactions, but who are provoked *more* often. Yet, as expected, children with these opposite change patterns did not differ in their overall change. Future work should examine how such change patterns are related to children's later adjustment, especially because overall measures might show treatment had little effect.

Our study was conducted in a summer program for at-risk youth, making it an interesting setting for studying change, but at the same time dictating caution. Although many of the children were referred from public schools, we have noted that they are elevated on standardized measures of aggression. Without a randomized design, we cannot make claims about program effectiveness. Our goal was not to evaluate this intervention, but to compare approaches to assessment that might be used in a wide range of treatment and school settings. Although we have no reason to believe that the change processes we found are unique to this summer program, additional research will be needed to assess the usefulness of our approach in other treatment and school settings and with other populations. Although our use of behavioral observations may have enhanced our ability to detect contextual variability in change, other research suggests that rating methods are also sensitive to children's social experiences (Farrell et al. 1998) and to the variability of behavior over situations (Dirks et al. 2007b). Future research will also be needed to assess the usefulness of these and other more efficient contextualized rating methods.

Mash and Hunsley (2005, p. 364) note that "blanket recommendations to use reliable and valid measures when evaluating treatments are tantamount to writing a recipe for baking hippopotamus cookies that begins with the instruction 'use one hippopotamus,' without directions for securing the main ingredient." Lacking directions, researchers often rely on overall ratings of problem behaviors, partly because their reliability has been "demonstrated", and partly because a premium is put on getting a simple summary of whether "improvement" occurred. Once an overall rating method is adopted, it is hard to critique its limited or mixed treatment of context, as this would suggest that some other instrument should have been used. Questions about what an "act" is, and how its frequency may be related to the surrounding context (Block 1989), distract from what is thought to be the main task of assessing behavior, personality, or change. The study of change ends up being stymied by its own reliance on reliable, but contextually uninformative, ratings of symptomatology (Dirks et al. 2007a; Eddy et al. 1998). If our findings do not provide directions for securing *the* main ingredient for assessing change, it is because multiple ingredients are needed. Even when reliable, overall behavior summaries potentially conflate changes in the individual's responses to social events and co-occurring changes in how often those events are encountered. Rather than viewing these complexities as distractions from the main task of evaluating change, we suggest that understanding them *is* the main task, much as it may be in the case of understanding behavior or personality in the first place.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington: University of Vermont.

August, G. J., Realmuto, G. M., Hektner, J. M., & Bloomquist, M. L. (2001). An integrated components preventive intervention for aggressive elementary school children: the Early Risers Program. *Journal of Consulting and Clinical Psychology, 69*(4), 614–626.

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.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science, 4*, 19–21.

Block, J. (1989). Critique of the act frequency approach to personality. *Journal of Personality and Social Psychology, 2*, 234–245.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park: Sage.

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

Chambers, J. M., & Hastie, T. J. (1992). *Statistical models in S*. Pacific Grove: Wadsworth.

Connor, D. F., Miller, K. P., Cunningham, J. A., & Melloni, R. H. (2002). What does getting better mean? Child improvement and measure of outcome in residential treatment. *The American Journal of Orthopsychiatry, 72*, 110–117.

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

Dirks, M. A., Treat, T. A., & Weersing, V. R. (2007a). Integrating theoretical, measurement, and intervention models of youth social competence. *Clinical Psychology Review, 27*, 327–347.

Dirks, M. A., Treat, T. A., & Weersing, V. R. (2007b). The situation specificity of youth responses to peer provocation. *Journal of Clinical Child and Adolescent Psychology, 36*, 621–628.

Dodge, K. A., McClaskey, C., & Feldman, E. (1985). Situational approach to the assessment of social competence in children. *Journal of Consulting and Clinical Psychology, 53*, 344–353.

Dodge, K. A., Dishion, T. J., & Lansford, J. E. (2006). *Deviant peer influences in programs for youth*. New York: Guilford.

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 and Adolescent Psychology, 37*, 314–326.

Eddy, J. M., Dishion, T. J., & Stoolmiller, M. (1998). The analysis of intervention change in children and families: methodological and conceptual issues embedded in intervention studies. *Journal of Abnormal Child Psychology, 26*, 53–69.

Farrell, A. D., Ampy, L. A., & Meyer, A. L. (1998). Identification and assessment of problematic interpersonal situations for urban adolescents. *Journal of Clinical Child Psychology, 27*, 293–305.

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.

Funder, D. C. (2009). Persons, behaviors and situations: an agenda for personality psychology in the postwar era. *Journal of Research in Personality, 43*, 120–126.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality, 23*, 369–401.

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

Gresham, F. M., & Elliot, S. N. (1990). *Social skills rating system manual*. Circle Pines: American Guidance Service.

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.

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 and Adolescent Psychology, 40*, 54–66.

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.

Hershberger, S. L., Plomin, R., & Pedersen, N. L. (1995). Traits and metatraits: their reliability, stability, and shared genetic influence. *Journal of Personality and Social Psychology, 69*, 673–685.

Hoffenaar, P. J., & Hoeksema, J. B. (2002). The structure of oppositionality: response dispositions and situational aspects. *Journal of Psychology and Psychiatry and Allied Health Disciplines, 43*, 375–385.

Kagan, J. (2001). The need for new constructs. *Psychological Inquiry, 12*, 84–103.

Kazdin, A. E. (2006). Arbitrary metrics: implications for identifying evidence-based treatments. *The American Psychologist, 61*, 42–49.

Kempes, M., Matthys, W., de Vries, H., & van Engeland, H. (2010). Children's aggressive responses to neutral peer behavior: a form of unprovoked reactive aggression. *Psychiatry Research, 176*, 219–223.

Ladd, G. W., & Profilet, S. M. (1996). The Child Behavior Scale: a teacher-report measure of young children's aggressive, withdrawn, and prosocial behaviors. *Developmental Psychology, 32*, 1008–1024.

Lewis, M. (2001). Issues in the study of personality development. *Psychological Inquiry, 12*, 67–83.

Lutz, M. N., Fantuzzo, J., & McDermott, P. (2002). Multidimensional assessment of emotional and behavioral adjustment problems of low-income preschool children: development and initial validation. *Early Childhood Research Quarterly, 17*, 338–355.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*, 362–379.

Matthys, W., Maassen, G. H., Cuperus, J. M., & Van Engeland, H. (2001). The assessment of the situational specificity of children's problem behaviour in peer-peer context. *Journal of Child Psychology and Psychiatry, 42*, 413–420.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont: Wadsworth.

McCrae, R. R., & Costa, P. T., Jr. (1999). A Five-Factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). New York: Guilford.

McDermott, P. A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. *Psychological Assessment, 5*, 413–424.

Measelle, J. R., John, O. P., Ablow, J. C., Cowan, P. A., & Cowan, C. P. (2005). Can children provide coherent, stable, and valid self-reports on the Big Five dimensions? A longitudinal study from ages 5 to 7. *Journal of Personality and Social Psychology, 89*, 90–106.

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

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.

National Institute of Child Health and Human Development Early Child Care Research Network. (2003). Does amount of time spent in child care predict socioemotional adjustment during the transition to kindergarten? *Child Development, 74*, 976–1005.

Nock, M. K., & Kurtz, S. M. S. (2005). Direct behavioral observation in school settings: bringing science to practice. *Cognitive and Behavioral Practice, 12*, 359–370.

Noftle, E. E., & Fleeson, W. (2010). Age differences in big five behavior averages and variabilities across the adult life span: moving beyond retrospective, global summary accounts of personality. *Psychology and Aging, 25*, 95–107.

Reynolds, C. R., & Kamphaus, R. W. (2002). *Behavior assessment system for children* (2nd ed.). Circle Pines: American Guidance Service.

Roberts, B. W., & Caspi, A. (2001). Personality development and the person-situation debate: it's déjà vu all over again. *Psychological Inquiry, 12*, 104–109.

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, 64*, 1177–1191.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology, 67*, 674–687.

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.

Tremblay, R. E. (2000). The development of aggressive behavior during childhood: what have we learned in the past century? *International Journal of Behavioral Development, 24*, 129–141.

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.

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: a review of seven coding schemes. *School Psychology Review, 34*, 454–474.

Walker, H. M., & McConnell, S. R. (1995). *Walker-McConnell scale of social competence and school adjustment, adolescent version*. NY: Wadsworth.

Weisz, J. R., Doss, A. J., & Hawley, K. M. (2005). Youth psychotherapy outcome research: a review and critique of the evidence base. *Annual Review of Psychology, 56*, 337–363.

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. (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.

Wright, J. C., Zakriski, A. L., & Drinkwater, M. (1999). Developmental psychopathology and the reciprocal patterning of behavior and environment: distinctive situational and behavioral signatures of internalizing, externalizing, and mixed-syndrome children. *Journal of Consulting and Clinical Psychology, 67*, 95–107.

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.

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.

Zakriski, A. L., Wright, J. C., & Cardoos, S. (2011). *Evaluating deviancy training within residential treatment: Individual and group effects of peer-nominated deviant talk*. Manuscript under review.