J Psychopathol Behav Assess
DOI 10.1007/s10862-013-9401-2

The Influence of Clinical Experience and Assessment Method on the Evaluation of Child Behavior Change

Anselma G. Hartley · Jack C. Wright ·
Audrey L. Zakriski · Catherine McCarthy

Abstract This research examined how people's ability to detect behavior change in simulated child targets is affected by their clinical experience and the assessment method they use. When using summary assessment methods that are widely employed in research and clinical practice, both inexperienced and experienced clinical staff detected changes in the overall frequency of targets' aggressive behavior, but were not uniquely influenced by changes in targets' reactions to social events. When using contextualized assessment methods that focused on conditional reactions, experienced staff showed greater sensitivity than novices to context-specific changes in targets' aggressive and prosocial reactions to aversive events. Experienced staff also showed greater sensitivity to context-specific changes in their overall impressions of change, but only for aggression. The findings show how clinically experienced judges become more attuned to *if…then…* contingencies in children's social behavior, and how summary assessment methods may hamper the detection of change processes.

Keywords Clinical experience · Child behavior change · Assessment · Social context · Social perception

Researchers and clinicians often advocate for greater attention to the contextual variability of behavior to deepen our understanding of children's psychosocial functioning and how it changes over time (Dirks et al. 2007; Kempes et al. 2010). Yet the tools that are often used to assess behavior and personality aggregate over contexts and emphasize overall behavioral tendencies (Cervone et al. 2001). Some researchers have suggested that summary methods lead raters to focus on overall behavior rather than how that behavior is influenced by contexts (Schwarz and Oyserman 2011). Others have argued that summary measures make it difficult to distinguish changes in people's social environments from changes in how they respond to events (Smith et al. 2009). Few studies, however, have tested these claims under controlled laboratory conditions (e.g., Hartley et al. 2013; Wright et al. 2001), and even fewer have done so using experienced raters (Wright et al. 2001). Using a methodology in which simulated child targets show distinct patterns of change over time in their environments and their behavioral responses, we examine the types of change novices and clinically experienced staff can detect and how summary assessment methods affect their sensitivity to change.

Summary assessments (e.g., Teacher Report Form and Child Behavior Checklist [TRF/CBCL], Achenbach and Rescorla 2001; Behavior Assessment System for Children, second edition [BASC-2], Reynolds and Kamphaus 2002) are often used to assess child behavior and how it changes. Such measures typically ask respondents to report how often a child displays various behaviors, leaving it to them to judge how situational factors should be taken into account. For example, a respondent might interpret "teases a lot" as a statement about the overall frequency of that act (Buss and Craik 1983), or the respondent might "implicitly contextualize" (Tellegen 1991; Wood and Roberts 2006) and conclude that the statement refers to relevant conditions (e.g., "teases *when provoked*"). Moreover, some items in these measures explicitly refer to contexts (e.g., "argues *when denied own way*"). Regardless of whether items are implicitly or explicitly contextualized,

A. G. Hartley (✉) · J. C. Wright · C. McCarthy
Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, 190 Thayer Street, Box 1821, Providence, RI 02912, USA
e-mail: anselma_hartley@alumni.brown.edu

J. C. Wright
e-mail: Jack_Wright@brown.edu

A. L. Zakriski
Department of Psychology, Connecticut College, Box 5646, 270 Mohegan Avenue, New London, CT 06320, USA
e-mail: alzak@conncoll.edu

Published online: 05 January 2014

variability over contexts is removed when summary scores are formed (e.g., aggression). Even if raters understand how behavior is influenced by situations, and even if this ability increases with experience, summary measures may capture little of this knowledge.

Advocates of functional behavioral assessment (Haynes et al. 2009) and social-cognitive learning approaches (Cervone et al. 2001) suggest that an analysis of behavior in context is needed to understand individual differences. Related research focuses on *if…then…* relationships between eliciting events and how people respond to them (e.g., *if* provoked, *then* aggressive) (Vansteelandt and Van Mechlen 1998). This approach proposes that individual differences are better revealed by the patterning of reactions rather than by behavior on average. For example, one child may act aggressively when teased, but may be teased rarely; another may be unlikely to act aggressively when teased, but may be teased often. Even though their responses to events differ, laboratory research with undergraduates and field research with teachers has shown that summary measures do not distinguish between such children because their overall rates of aggression are the same (Wright et al. 2001). Similarly, field research has found that staff ratings of overall aggression do not distinguish between two types of functionally distinct children: those who became more likely over time to experience aversive events (e.g., peer tease) but became less likely to react aggressively when this occurred, versus those who showed the opposite changes in these events and reactions (Wright et al. 2011).

Numerous studies demonstrate that observers are sensitive to contextualized behavior patterns (Fournier et al. 2008; Kammrath et al. 2005), and incorporate situational information into their personality judgments (Smith and Collins 2009). However, these sensitivities depend on the conditions under which raters process information. People have difficulty integrating situational influences when cognitive load is high, and when salience of the stimuli is low (Chun et al. 2002; Gilbert and Malone 1995). Sensitivity to behavioral and situational information also depends on raters' affective state (Hunsinger et al. 2012), the format of the assessment instrument they are asked to complete (Schwarz and Oyserman 2011), and their statistical knowledge and investment in the social target (Schaller 1992). Thus, even if people have the capacity to detect *if…then…* contingencies between contexts and behaviors and use them to evaluate change, it is possible that summary assessment methods divert their attention away from these contingencies and lead them to focus on overall behavior. Little is known about whether clinical experience counteracts this.

A large body of research has raised questions about the value of clinical experience (Dawes 1994; Garb 1989, 1998), and has shown that clinical predictions made by professionals are often inferior to actuarial methods that are purely statistical and do not involve complex human judgment (Dawes et al. 1989; Grove 2005; Grove et al. 2000). This raises the possibility that clinical experience has little effect on how raters disentangle situational and behavioral information when assessing behavior and how it changes, and that raters of all types have equal difficulty with these tasks. Yet, other research in both clinical and non-clinical domains indicates that experts and novices differ in their judgment abilities, if sometimes modestly and in task-specific ways. Research in non-clinical domains (Garcia-Retamero and Dhami 2009) shows that experts can form quicker, more accurate judgments by using valid cues from the environment, especially familiar ones (Ericsson and Lehmann 1996; Hertwig et al. 1999; Kahneman and Klein 2009). In the clinical domain, experts are more sensitive than novices to context-behavior covariation (Dawson et al. 1989), and are better able to apply statistical heuristics if it is apparent that statistical reasoning is appropriate (Nisbett et al. 1983). Experienced counselors differ from novices in several respects, including better short- and long-term memory for domain-specific information and richer cognitive schemata about case materials (see Spengler et al. 2009). This suggests that clinically experienced raters may outperform novices when social stimuli are representative of the situations and behaviors they often observe, when the task involves little inference or extrapolation, and when the assessment methods used "optimize the expression of [their] expertise" (Westen and Weinberger 2005, p. 1257). Thus, despite well-documented shortcomings of clinical judgment, it remains possible that more experienced clinicians can detect and track key *if…then…* contingencies in children's behaviors, as long as they are given tools that facilitate this level of processing.

To address these issues, we build on past research on contextual assessment and incorporate the dimension of clinical experience. Using an experimental paradigm developed in studies of undergraduate raters (Hartley et al. 2013), we examine differences between novice and more experienced clinical staff in their ability to assess behavior change. We created targets that changed over time, both in how often they encountered aversive social events ("event rates") and in the conditional probability of their aggressive reactions when these events occurred ("reaction rates"). We focused on aversive peer and adult situations because these demanding events and maladaptive reactions to them have been widely studied in research on children's behavior problems (see Skinner and Zimmer-Gembeck 2007). Two of the targets we used were "simple" in that the rates of events they encountered and the rates of their reactions to those events changed in the same direction. One of these targets experienced an increase in how often he was provoked by peers and disciplined by adults and also showed an increase in the conditional probability of his aggressive reactions to these events. For the other, both the event rate and reaction rate decreased. For these targets, the combined effect of changes in their environments and in their reactions was that their overall rate of aggressive behavior either increased or decreased.

The two other targets we used were "complex" in that their event rate and reaction rate showed opposing changes over time. One target experienced an increase in the rate of aversive events but became less likely to react aggressively to those events. For this target, the combined effect of these offsetting changes resulted in no change in his overall rate of aggressive behavior. The other complex target experienced a decrease over time in aversive events but became more likely to react aggressively when those events occurred. Although this target is the functional opposite of his counterpart, he was equivalent in that he also showed no change over time in his overall rate of aggressive behavior. These complex targets are of special interest because they exhibit clinically meaningful behavior change that is not evident in their overall frequencies.

To clarify how raters' sensitivity to change is influenced by the way behavior is assessed, we administered scales from two popular summary instruments often used to measure child behavior (TRF, Achenbach and Rescorla 2001; BASC-2, Reynolds and Kamphaus 2002) as well as summary items directly matched to the behaviors manipulated in the vignettes. We also asked participants to rate how often targets encountered aversive and nonaversive social events, and how often targets responded with aggressive or prosocial behaviors when these events occurred. Finally, we asked participants to provide their overall impressions of change in targets' behavior.

We tested three main predictions. First, we expected summary measures to focus raters' attention on the child's behavior rather than on the surrounding social situation, thereby leading them to be primarily sensitive to targets' overall behavior frequencies. Specifically, participants' ratings using summary scales should distinguish between the two simple targets, as their overall frequencies of aggression changed. However, these measures should not distinguish between the two complex targets who showed no overall change despite opposite changes in their aggressive reactions. Because summary assessments restrict opportunities to report context-specific knowledge in the first place and then filter it out in the aggregation process, we expected novice and experienced judges to show similar results on these measures.

Second, we expected all participants to show greater sensitivity to targets' context-specific reactions when the judgment task focused their attention on *if…then…* contingencies between contexts and behaviors. We also expected more experienced judges to be more sensitive than novices to changes in targets' reactions, and to distinguish better between changes in targets' reactions and changes in their environments. Therefore, even though experienced and novice judges should perform comparably when using summary scales, experienced judges should display greater sensitivity when using a contextualized judgment task. Such evidence could help clarify how and why novice and experienced staff differ in their conclusions about behavior change, and how potentially important changes can be missed when only summary measures are used.

Third, we expected experienced judges' overall impressions of change to be more influenced by changes in targets' reactions to events than by changes in targets' social environments. We predicted novices' overall impressions of change would be based more on changes in the overall frequency of targets' behaviors. Again, the complex targets provided a critical test of this hypothesis, since their reactions changed even though their overall behavior frequencies did not.

Method

Participants

Participants were 141 (67 male, 74 female) staff working at a short-term (6-week) residential summer treatment program for at-risk youth located in New England (M_{age}=22.33, SD= 2.58). Participants had worked in the agency for 1–9 summers (M=1.96, SD=1.71). Their educational background was as follows: 73 (52 %) were in college (3 sophomores, 22 juniors, 46 seniors); 51 (36 %) had a bachelor's degree; and 17 (12 %) had a master's degree. Their major areas of study (for current or highest degree) were: psychological sciences and education (69 %), humanities (7 %), social sciences (4 %), life sciences (3 %), and undeclared (17 %).

The index of experience we used was based on number of hours of experience in the agency, derived from their employment history. The agency states that staff receive 750 supervised internship hours per summer. Supervisors have a longer session and additional responsibilities, including admission interviews, conducting therapy, supervision meetings, writing clinical reports, and planning time during orientation and reorientation weeks. Based on consultation with administrators, we estimated this as an additional 25 % in hours of experience (188 h) for each summer as a supervisor. The final experience index ranged from 750 to 7,500 h (M=1513.30, SD= 1385.11). We also examined a second index, which summed hours in the agency and participants' estimates of hours of clinical experience in other settings (M=1995.71, SD= 1619.90). These two indices were highly correlated (r=.98, p<.001). We report analyses using the first index because it is a conservative measure and was unaffected by bias in participants' self reports.

Experimental Stimuli

The stimuli were based on Hartley et al. (2013), which in turn were based on Wright et al. (2001), except that they showed the target's behavior at two time points rather than one. Participants viewed timed PowerPoint® slides describing a target (a fictitious 11-year-old boy named Dan) in a 6-week residential summer program. Thirty-two slides described the target's interactions at the beginning of the summer program (Time 1); 32 slides described the target's behavior at the end of the program (Time 2). Four targets were created. One encountered an increase in aversive events and showed an increase in aggressive reactions to those events (R+/E+) ("+" = increase). The second showed a decrease in both event rate and reaction rate (R−/E−) ("−" = decrease). The third encountered an increase in aversive events, but showed a decrease in his aggressive reactions (R−/E+). The fourth had the reverse arrangement (R+/E−).

Each slide described the setting and an interaction between Dan and another person. The setting, agent, agent action, target name, and target's reaction appeared in the same order. "Aversive events" consisted of two types of peer events (tease, argue) and two types of adult events (warn, discipline). "Nonaversive events" consisted of two types of peer events (prosocial talk, ask) and two types of adult events (prosocial talk, ask/instruct). Target reactions were either aggressive (tease, argue) or nonaggressive (prosocial talk, ask). An example of an aversive peer event with an aggressive reaction is: "In basketball, a boy says, 'It's my turn to play point guard. You can't even make a lay-up!' Dan says, 'No way. I was here first, so get lost.'"

Table 1 shows the probabilities of aversive events, p(E), conditional probabilities of aggressive reactions to those events, p(R | E), the probabilities of overall aggression, p(R), and the frequencies of the stimuli for each condition. The probability of aversive events is obtained by dividing the number of aversive events per time point by the total number of vignettes per time point (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 (note that aggressive behaviors occurred only in response to aversive events). The overall probability or "base rate" of aggression, p(R) is obtained by p(E) * p(R | E); this is equivalent to the number of aggressive behaviors per time divided by the total number of vignettes per time. The "simple" R+/E+ and R−/E− 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 time (rows 1 and 4). The "complex" R+/E− and R−/E+ targets (rows 2 and 3) differed in the conditional probability of their aggressive reactions to aversive events, but had equal base rates of aggression at each time and therefore no change in overall aggression.

Dependent Measures

Open-Ended Description Participants were first instructed to describe briefly in writing "what a new staff member about to start working with Dan would most need to know in order to understand him and work effectively with him." Because these responses were brief, and because our main focus was on structured measures, these descriptions are not reported here.

Modified Teacher Report Form (TRF) As in Wright et al. (2001), we used a subset of the 118 items from the 2001 version of the TRF (Achenbach and Rescorla 2001) to avoid fatiguing participants. Specifically, we used the scale relevant to this study (aggression, 20 items) and a contrast scale (withdrawal, 8 items). The word "teachers" was changed to "staff" (to match stimuli) for one item. Items were rated using the TRF's 0–2 scale. Only the aggression scale was analyzed. We have noted that some items in summary measures provide some contextualization: for aggression, there were three such items (destroys his own things; destroys property belonging to others; defiant, talks back to staff) and the others were acontextual (e.g., refuses to talk). Internal consistency for the TRF aggression scale (Cronbach's α = .90) was comparable to that presented in Achenbach and Rescorla (2001); internal consistency was lower for the TRF withdrawal scale

Table 1 Properties of the four experimental targets over time

	Time 1			Time 2		
Condition	p(E)	p(R|E)	p(R)	p(E)	p(R|E)	p(R)
R−/E−	0.75 (24/32)	0.75 (18/24)	0.56 (18/32)	0.25 (8/32)	0.25 (2/8)	0.06 (2/32)
R−/E+	0.25 (8/32)	0.75 (6/8)	0.19 (6/32)	0.75 (24/32)	0.25 (6/24)	0.19 (6/32)
R+/E−	0.75 (24/32)	0.25 (6/24)	0.19 (6/32)	0.25 (8/32)	0.75 (6/8)	0.19 (6/32)
R+/E+	0.25 (8/32)	0.25 (2/8)	0.06 (2/32)	0.75 (24/32)	0.75 (18/24)	0.56 (18/32)

Note that p(R) = p(E) * p(R|E). "+" indicates increase; "–" indicates decrease in event or reaction rate. *E* event, *R* reaction. Frequencies on which probabilities and conditional probabilities were based are in parentheses; for p(E) and p(R), the denominator is always the total number of vignettes per time (32), and for p(R|E), the denominator is the number of aversive events per time

p(E) probability of aversive event, *p(R|E)* probability of aggressive reaction to aversive event, *p(R)* base-rate probability of aggressive behavior

(α=.68), likely because this behavior was not shown in the experimental stimuli.

Modified Behavior Assessment System for Children (BASC-2) Because the TRF does not assess prosocial behaviors in its main scales, we also included the social skills scale from the BASC-2 (Reynolds and Kamphaus 2002). This scale consists of 11 items, including several that assess prosocial behaviors (e.g., "compliments others"); henceforth we refer to this scale as "prosocial" for brevity. Only one item was contextualized ("congratulates others when good things happen"). Items were rated on a 0–2 scale. Internal consistency for the BASC prosocial scale (α=.89) was similar to that reported in Reynolds and Kamphaus (2002). For efficiency of administration, items from this scale were presented together with those from the TRF on one page. Items were ordered based on how they appear on the BASC and TRF. The TRF aggression and BASC prosocial scales were moderately negatively correlated (r=−.56, p<.001).

Behavior, Event, and Reaction Measures To clarify whether participants detected the manipulated event rates and reaction rates at each time point, additional items focused on the specific events and behaviors that were shown in the stimuli. All of the following items were rated on a 6-point scale (0="never", 5="almost always"). Participants first rated the overall frequency of Dan's aggressive and prosocial behaviors shown during Time 1 using four items ("Dan teased, threatened, or bossed"; "argued or quarreled"; "talked politely or made friendly requests"; "behaved in a polite or friendly way"). They then rated how often Dan encountered aversive and non-aversive events at Time 1, using four items ("peers teased, threatened, or bossed Dan"; "peers talked politely or made friendly requests"; "adults warned or disciplined"; "adults complimented or made friendly requests"). Next, they rated Dan's reactions when a given event occurred, using 16 items (4 events × 4 reactions). Participants read, "Indicate how often Dan showed each reaction to the event described." After each event prompt (e.g., "When 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"); reactions were the same as the behaviors noted above. Participants then rated behaviors, events, and reactions shown during Time 2. Because the overall aggressive and prosocial behavior items were similar to and correlated with the TRF aggression and BASC prosocial scales (rs=.91, .80, respectively, ps<.001), these additional summary measures are not presented here. The aversive event scale was scale was moderately correlated with the TRF aggression scale and negatively correlated with the BASC prosocial scale (rs=.47, −.46, respectively, ps<.001); the aggressive reaction scale showed the same pattern (rs=.68, −.46, respectively, ps<.001); and the prosocial reaction scale was negatively correlated with the TRF aggression scale and moderately correlated with the BASC prosocial scale (rs=−.57, .40, respectively, ps<.001).

Impressions of Change Nine items assessed participants' overall impressions of change using 7-point scales, including: increases/decreases in Dan's aggressive, withdrawn, and prosocial behavior; improvement/worsening in his behavior toward peers and adults; and changes in the aversiveness/supportiveness of peers' and adults' behavior toward him. Although these 9 items showed very high internal consistency (α=.94), we focus on ratings of aggression change and prosocial change (1="increased greatly", 4="no change", 7="decreased greatly"), as these behaviors were most relevant to our stimuli. We also analyzed overall target change by aggregating all 6 items that asked about the target's behavior change (α=.91). To facilitate graphical comparisons across measures, we reversed the scale so that "7" indicated an increase and "1" a decrease; this had no effect on any findings we report.

Ordering of Assessments Dependent measures were administered in the order just presented, with the open-ended measure first at each time point. Rather than counterbalancing the formal assessments that followed at each time point, the TRF and BASC scales were always administered first, before the measure that drew attention to situation-behavior contingencies. We did this to preserve the integrity of the summary measures and to resemble as closely as possible how summary measures would be completed in a stand-alone assessment. Finally, participants' impressions of change were collected once, after all other measures at the second time point.

Procedure

Participants were run in groups of 5–10 on separate computers, and were randomly assigned to condition, to which the experimenter was blind (R−/E−, n=37; R−/E+, n=35; R+/E−, n=33; R+/E+, n=36). They were told they would read about "Dan," an 11-year-old boy at a residential summer program. Using the measures just described, they completed the following steps: (1) read vignettes for Time 1, each for 9 s; (2) provided open-ended descriptions, TRF/BASC ratings, and ratings of overall behavior, events, and reactions; (3) read vignettes for Time 2; (4) repeated Step 2; (5) rated their impressions of change. Participants reported their past experience and training in a separate session following the experiment.

Results

We performed generalized estimating equations (GEE) to model participants' ratings on the repeated dependent measures because such methods take correlations among those measures into account when estimating standard errors (Liang and Zeger 1986). These analyses were run in R (R Development Core Team 2012; Halekoh et al. 2006). All dependent measures were continuous; thus, models assumed a normal distribution with an identity link function and were specified within the Gaussian family. Following Hox's (2010) recommendations for centering predictors, time, reaction condition, event condition, and experience were grand mean centered. The experience predictor was log-transformed due to its positive skew, then grand-mean centered. We restricted the model to include terms of theoretical interest: main effects for time, event condition, and reaction condition, two-way interactions between event, reaction, and time, three-way interactions between event or reaction with time and experience, and the four-way interaction. We initially included a term for time × event × reaction, but this term did not make a contribution and was not central to our predictions about experience, so it was dropped. GEE analyses were run to predict participants' Time 1 and 2 TRF, BASC, aggressive reaction, prosocial reaction, and aversive event ratings. For graphical illustration, we generated predicted values for "low" and "high" experience as 1 *SD* below or above the mean for experience. Impressions of overall change were assessed at a single time point (i.e., end of the experiment), and were therefore analyzed using analysis of covariance (ANCOVA) with experience as the covariate. Because there were fewer terms, and because the reaction × event interaction was of interest concerning impressions of change, the full unrestricted ANCOVA model was used.

Summary Assessment If participants' TRF ratings were influenced by changes in the simple targets, but not by changes in the complex targets who diverged in their aggressive reactions, we would expect to see comparable effects for the time × event and time × reaction interaction terms. As shown in Table 2, these time × reaction and time × event interactions were significant. As shown in Fig. 1a–b, participants at all levels of experience distinguished between the simple targets whose overall behavior rates changed (R+/E+, R−/E−; black lines). However, they did not distinguish between the complex targets whose event and reaction changes conflicted (R+/E−, R−/E+; dashed lines); they rated these targets as unchanged, as one would expect based on the fact that the manipulated overall rates of aggression for these targets did not change over time. We did not expect these results to be moderated by participants' level of experience, and thus did not predict 3-way interactions with experience. As shown in Table 2, no such 3-way interactions were found for the TRF aggression ratings. No other results were expected or found.

Parallel analyses were performed for participants' BASC prosocial ratings. Results were similar to those obtained for TRF aggression (see Table 2 and Fig. 1c–d) except the pattern of findings was inverted because aggression was the manipulated variable (e.g., "R−" means aggressive reactions decreased). The time × reaction and the time × event terms were significant. Participants clearly judged the simple targets (R+/E+, R−/E−) as showing increased or decreased overall prosocial behavior. For complex targets, although participants rated the R−/E + target as unchanged, their prosocial ratings of the R+/E− target were influenced more than expected by decreases in aversive events. Participants rated this target as becoming more prosocial even though he became *more aggressive* (and less prosocial) to aversive events. Time was also an unexpected significant predictor, with participants on average reporting prosocial behavior as increasing. As expected, experience effects were not revealed on this summary prosocial measure. Overall, when using summary measures, participants did not distinguish between complex targets who diverged in the direction of their reaction change, regardless of experience. Instead, all participants' ratings tracked targets' overall behavior rates.

Reaction Ratings We predicted that experienced participants would show greater sensitivity to changes in *if…then…* reactions when explicitly contextualized measures were used. For aggressive reactions, the time × reaction × experience term was significant, indicating that participants with greater experience showed more sensitivity to aggressive reactions than those with less experience (see Table 3 and Fig. 1e–f). For aggressive reaction ratings, novice participants did not distinguish between the complex targets, whereas those with more experience did (dashed lines; Fig. 1e–f). As expected, the time × reaction term was significant, indicating that aggressive reaction ratings were influenced by manipulated changes in aggressive reactions for all participants. Table 2 also shows that the time × event term was significant, indicating that event rates also influenced participants' aggressive reaction ratings.

Results for prosocial reactions to aversive events were similar to those for aggressive reactions, except the overall pattern was again inverted (see Table 3 and Fig. 1g–h). As predicted, the time × reaction × experience interaction was significant; the time × event × experience term also significantly contributed. Participants with greater clinical experience were more sensitive to changes in the target's prosocial reactions to aversive events than were those with low experience. The time × reaction and time × event interactions were significant; participants' ratings of prosocial reactions were somewhat influenced by the aversive events that the target encountered.

Event Ratings We expected all participants, regardless of experience, to be sensitive to the rates of aversive events. As expected, we found that the time × event interaction was significant (Table 4), indicating that participants detected changes in how often aversive events occurred. However, there was also an unexpected, smaller time × reaction

Table 2 Summary of GEE models predicting Teacher Report Form (TRF) and Behavior Assessment System for Children (BASC) ratings from time, reaction condition, event condition, and experience

	TRF				BASC			
Predictors	χ^2	Est.	SE	95 % CI	χ^2	Est.	SE	95 % CI
Time	2.48	−0.60	0.38	[−1.35, 0.15]	9.90	1.00**	0.31	[0.39, 1.61]
Reaction (R)	0.00	−0.03	0.77	[−1.55, 1.49]	0.47	−0.48	0.71	[−1.89, 0.93]
Event (E)	0.00	0.00	0.77	[−1.52, 1.52]	1.85	0.96	0.70	[−0.43, 2.35]
Time × R	166.50	9.68***	0.75	[8.20, 11.17]	60.75	−4.90***	0.63	[−6.15, 3.65]
Time × E	183.57	10.20***	0.75	[8.72, 11.69]	134.92	−7.31***	0.63	[−8.56, −6.06]
Time × R × Exp	0.04	0.20	1.06	[−1.90, 2.29]	0.94	0.89	0.92	[0.93, 2.71]
Time × E × Exp	0.08	−0.28	1.06	[−2.38, 1.82]	2.84	1.53	0.91	[−0.27, 3.33]
Time × R × E × Exp	1.49	2.53	2.14	[−1.71, 6.77]	0.12	1.81	1.87	[−1.89, 5.51]

χ^2 Wald's chi-square, *Est.* estimate, *Exp* experience

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

interaction, showing that the target's aggressive reaction rates influenced participants' event ratings. Reaction was also a significant predictor. Both experienced and inexperienced participants rated targets whose aggressive reactions increased (R+/E−; R+/E+) as encountering more aversive events than those whose aggressive reactions decreased (R−/E+; R−/E−). We return to this finding in the Discussion.

Impressions of Change We predicted participants with greater experience to distinguish between complex targets based on how their aggressive *reactions* changed, whereas less experienced participants would base their impressions on *overall* aggression change. There were main effects for reaction and event conditions, $Fs(1, 133)=138.95, 74.03$, $ps<.001$, as well as the expected reaction × experience interaction, $F(1, 133)=4.81$, $p<.05$. As predicted, participants with low experience were sensitive primarily to changes in the targets' base-rates of aggression whereas those with higher experience were sensitive to how targets' aggressive reactions changed (see Fig. 2a–b). Only experienced raters' impressions of aggression change distinguished between the two complex targets.

Targets displayed aggressive behavior only in response to aversive events. In contrast, prosocial behavior was sometimes displayed in response to aversive events, but was also displayed to *all* nonaversive events. If raters tracked both of these prosocial reaction patterns, their impressions of changes could more closely track behavior base-rates. There were main effects for reaction and event conditions, $Fs(1, 133)=29.86, 38.88$, $ps<.001$, and no effects of experience: participants judged change in targets' prosocial behavior based on changes in the overall rate of prosocial behavior, not prosocial reactions. A reaction × event effect, $F(1, 133)=6.85$, $p<.01$, indicated that the target who increased in both aggressive reactions and aversive events (R+/E+) was rated as showing the greatest decrease in prosocial behavior (Fig. 2c–d). No other significant effects were expected or found. We also tested overall impressions of behavior problem change, and found significant main effects for reaction and event condition, $Fs(1, 133)=177.44, 141.12$, $ps<.001$. As with prosocial change, the event × reaction interaction was also significant, $F(1, 133)=10.87$ $p<.01$. The simple targets were seen as most clearly increasing and decreasing in their total problem behavior (Fig. 2e–f). No other significant effects were expected or found.

Discussion

This study used an experimental paradigm to examine how clinical experience and assessment method affect people's ability to report on distinct situation-specific patterns of change. Three major findings emerged. First, when using summary measures of the type commonly employed in research and clinical practice, all staff provided ratings that tracked overall behavior frequencies. Clinical experience did not influence their ability to distinguish between complex targets who showed opposite changes in their aggressive reactions to aversive events. Second, when explicitly asked about rates of targets' aggressive and prosocial reactions to aversive events, clinically experienced raters showed more sensitivity to these event-behavior relationships and better distinguished between the complex targets. Third, when asked to provide their impressions of aggression change, more experienced staff were clearly influenced by changes in targets' aggressive reactions, whereas less experienced staff were influenced mainly by changes in the overall frequency of targets' aggressive behavior. These findings suggest that with greater clinical experience, judges become more sensitive to

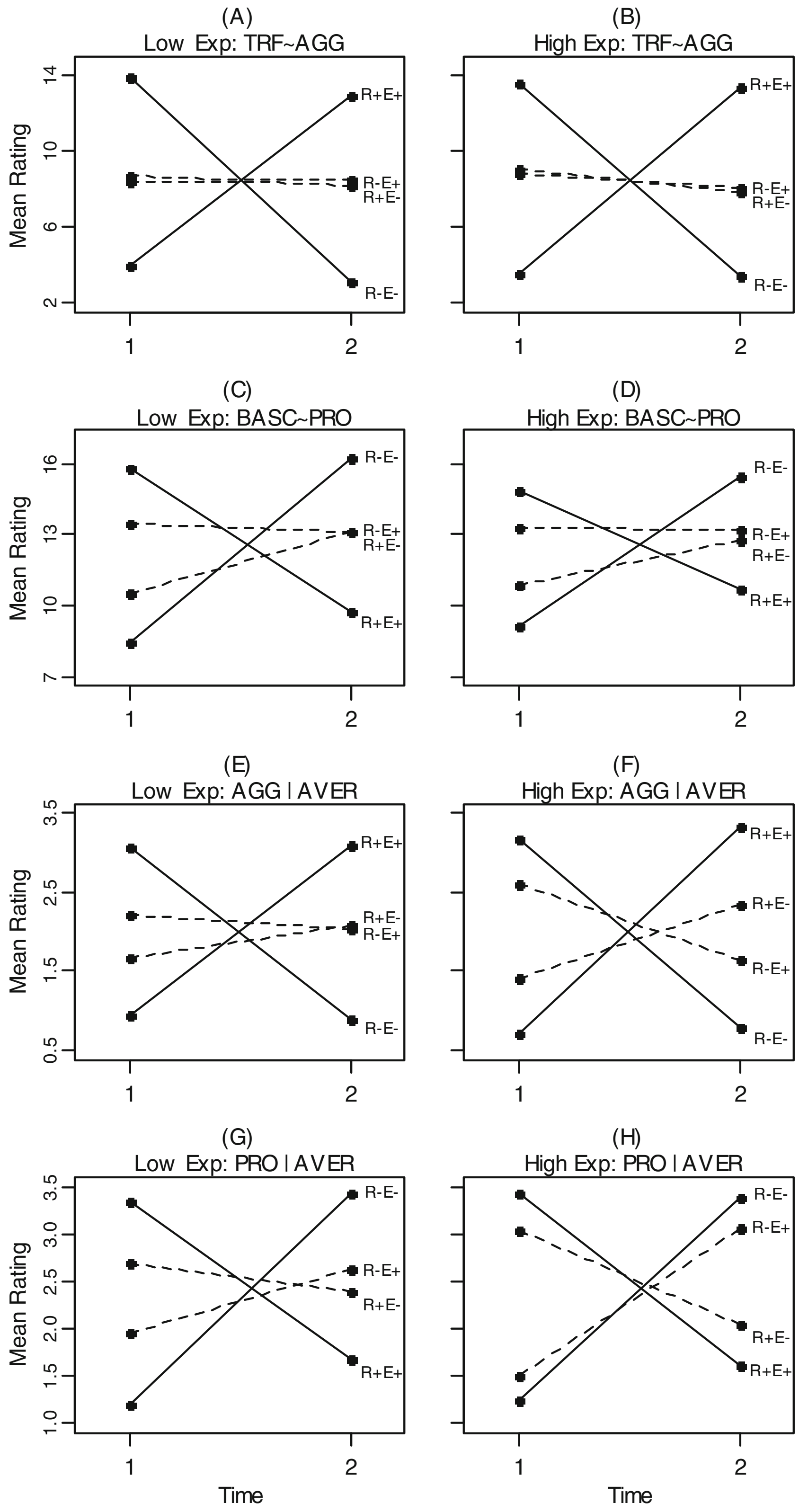

Fig. 1 GEE model predicted values for Teacher Report Form (TRF) aggression, Behavior Assessment System for Children (BASC) prosocial, aggressive reaction, and prosocial reaction ratings by Time. *Left panels* are participants with low experience (−1 *SD* below the mean); *right panels* are participants with high experience (+1 *SD* above mean). *Solid lines* represent simple conditions; *dashed lines* represent complex conditions. *TRF ~ AGG* TRF aggression, *BASC ~ PRO* BASC prosocial, *AGG | AVER* aggressive reactions to aversive events, *PRO | AVER* prosocial reactions to aversive events

situation-specific patterns of child behavior. This sensitivity can be revealed both in situation-specific behavior ratings and in people's impressions of behavior change, but is masked by summary assessments.

Our findings do not support the view that participants "implicitly contextualize" summary measures so as to assess *if*…*then*… reaction patterns (see Wood and Roberts 2006). Rather, these measures detected overall behavior rates, and

Table 3 Summary of GEE models predicting aggressive and prosocial reaction ratings from time, reaction condition, event condition, and experience

	Aggressive reaction				Prosocial reaction			
Predictors	χ^2	Est.	SE	95 % CI	χ^2	Est.	SE	95 % CI
Time	0.33	0.05	0.09	[−1.13, 0.23]	4.62	0.22*	0.10	[0.02, 0.42]
Reaction (R)	0.81	−0.10	0.12	[−0.34, 0.14]	2.79	0.22	0.13	[−0.04, 0.48]
Event (E)	1.62	0.15	0.12	[−0.09, 0.39]	0.03	−0.02	0.13	[−0.28, 0.24]
Time × R	242.97	2.98***	0.19	[2.60, 3.36]	187.27	−2.84***	0.21	[−3.26,− 2.42]
Time × E	80.45	1.72***	0.19	[1.34, 2.10]	27.09	−1.09***	0.21	[−1.51, −0.67]
Time × R × Exp	7.01	0.77**	0.29	[0.20, 1.34]	3.99	−0.62*	0.31	[−1.23, −0.01]
Time × E × Exp	1.17	−0.32	0.29	[−0.89, 0.25]	5.42	0.68*	0.30	[0.09, 1.27]
Time × R × E × Exp	0.23	0.27	0.57	[−0.86, 1.40]	0.04	−0.13	0.61	[−1.34, 1.08]

χ^2 Wald's chi-square, *Est.* estimate, *Exp* experience

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

therefore missed changes in targets' situation-specific reactions. When using scales taken from the TRF and the BASC, neither inexperienced nor experienced staff could distinguish between the complex targets even though these targets showed diverging changes in their aggressive reactions to aversive events. These findings dictate caution when judging whether change scores based on summary measures indicate changes "within the person" (see Block 2010). Such scores may not discriminate between distinct processes that contribute to overall behavioral output, such as changes that occur in a child's social environment, changes in their reactions, or both.

In contrast to their performance on scales found in summary measures (TRF, BASC), when explicitly asked about aggressive and prosocial reactions, clinically experienced staff showed greater sensitivity to changes in targets' reactions than did less experienced staff. As predicted, more experienced staff distinguished between complex targets based on whether they increased or decreased in aggressive or prosocial reactions to aversive events. Explicitly asking these raters about targets' reactions apparently helped them access information about situation-behavior relationships and enabled them to demonstrate their expertise in ways that summary measures did not (Westen and Weinberger 2005). Similar but weaker effects of experience were seen in participants' impressions of aggression change. When asked to provide an overall judgment of how the target's aggression changed, more experienced staff relied on changes in the target's aggressive reactions. Less experienced staff relied more on changes in the target's overall aggression.

Table 4 Summary of GEE model predicting aversive event ratings from time, reaction condition, event condition, and experience

	Aversive event			
Predictors	χ^2	Est.	SE	95 % CI
Time	2.75	−0.11	0.07	[−0.25, 0.03]
Reaction (R)	9.42	0.28**	0.09	[−0.10, 0.46]
Event (E)	3.77	−0.18	0.09	[−0.36, 0]
Time × R	43.37	0.87***	0.13	[0.61, 1.13]
Time × E	228.35	2.01***	0.13	[1.75, 2.27]
Time × R × Exp	2.03	−0.30	0.22	[−0.74, 0.14]
Time × E × Exp	0.50	−0.15	0.21	[−0.57, 0.27]
Time × R × E × Exp	0.19	0.18	0.42	[−0.65, 1.01]

χ^2 Wald's chi-square, *Est*. estimate, *R* reaction condition, *E* event condition, *Exp* experience

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

Alternative interpretations of these results deserve scrutiny. One is that less experienced staff ignored the surrounding social situation and instead attended to targets' aggressive behavior because it was especially salient to them. This interpretation is unsatisfying, in part because participants at all levels of experience showed moderate sensitivity to how often targets encountered aversive events. A second interpretation is that, although both novices and more experienced participants attended to events, experienced judges were better able to encode and track *if…then…* situation-behavior contingencies shown in our stimuli. The stimuli in our study were designed to be representative of the situations and behaviors experienced staff often observe in their interactions with children. Clinical training may also include explicit instruction in monitoring and attending to children's situation-specific behavior patterns (see Haynes et al. 2009), thereby reinforcing the effects of repeated exposure. This interpretation is consistent with evidence that experts in other domains often focus on information that is most relevant to a task, whereas novices may be less selective and use less efficient judgment strategies (see Gigerenzer and Gaissmaier 2011).

Although experienced judges' familiarity with situation-behavior relationships may have contributed to their performance, our results show that the assessment method also

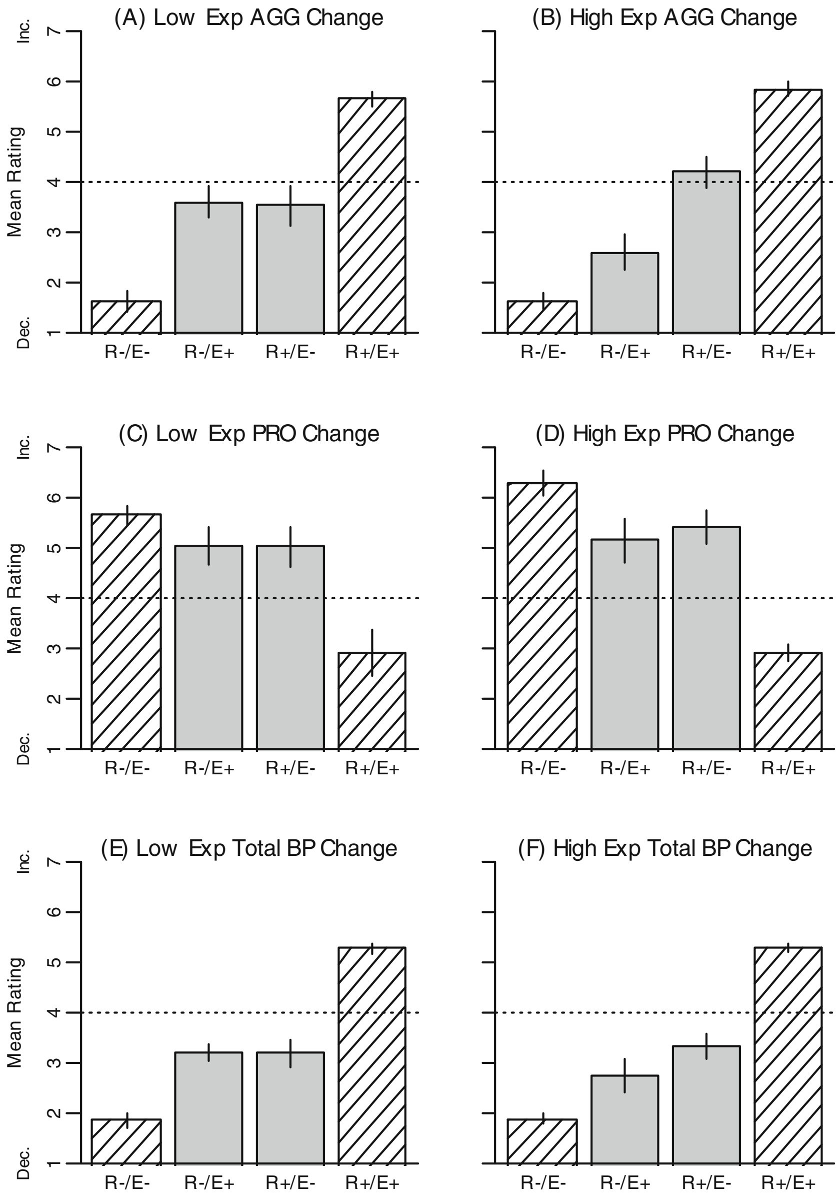

Fig. 2 Mean ratings for overall aggression, prosocial, and total change judgments by condition. *Top row* shows results for aggression (AGG) change judgments; *middle row* for prosocial (PRO) change judgments; *bottom row* is judgments of total target behavior problem change (Total BP). *Left panels* indicate participants with low experience (−1 *SD* below the mean); *right panels* indicate participants with high experience (+1 *SD* above mean). The y-axis shows mean change from decrease (Dec.) to increase (Inc.); 4=no change (*dotted line*). Experimental conditions are on the abscissa; *hatched bars* represent the simple conditions; *grey bars* represent the complex conditions. *Error bars* indicate +/− 1 SEM

influences whether they will express their sensitivity to these relationships (Westen and Weinberger 2005). Despite their differences on our conditional probability task, experienced and inexperienced staff performed similarly when they were asked to assess behavior change using scales from popular behavior checklists that do not emphasize situation-behavior contingencies. Experienced raters appeared able to operate either in a "summary" mode or a "conditional" mode, depending on how the assessment tool framed the task (Schwarz and Oyserman 2011). Other research has shown that nominal experts—including clinical psychologists and counselors—perform as poorly as novices on tasks that appear relevant to their expertise (see Garb 2006; Shanteau et al. 2002; Shanteau et al. 2002). Thus, in the present context, summary assessment methods, though relevant to research and clinical practice with children, do not appear to engage certain sensitivities that experienced staff possess. Future research in this area should examine both the conditions under which experts can take advantage of their domain-specific knowledge and heuristics as well as the conditions that interfere with their performance (see Chase and Simon 1973; Gigerenzer and Gaissmaier 2011).

Participants at all levels of experience detected changes in how often aversive events occurred, but their ratings of event rates were also influenced by the targets' aggressive reactions to events. It is possible that participants used aggressive reactions to events to disambiguate events that were not unambiguously aversive (Trope et al. 1988). Likewise, our analyses suggest that changes in aversive event rates affected raters' perceptions of behavior change, as measured by the BASC prosocial scale and by items that assessed targets' aggressive and prosocial reactions to events. This influence of events on ratings of targets' behavior was typically smaller than the influence of manipulated reaction rates, but it raises interesting questions about how raters may use clues in the environment (e.g., peers are teasing a child less over time) to infer that the child must also be changing (e.g., the target must be getting less aggressive). Future research should manipulate the degree of salience of events and reactions to clarify how people interpret changes in the surrounding environment when rating child behavior.

Our study has limitations that suggest future directions for research. The TRF and BASC are used more often in research and clinical practice than are context-sensitive measures like those used in this study. Moreover, summary measures may be the main assessment tool when efficiency is needed (Dever et al. 2012). We therefore presented the items from the TRF and BASC before the other dependent measures. One might argue this drew attention to targets' overall behaviors and led novice staff to conclude that their ratings on the reaction measures should be consistent with their (initial) summary ratings, whereas experienced staff showed greater flexibility over these tasks. This interpretation is consistent with the view that expertise is multifactorial, including domain-relevant knowledge and appropriate decision heuristics, but also so-called "psychological characteristics" such as self-confidence and adaptability to new tasks (Shanteau et al. 2002). Future research should clarify whether experts show flexibility only when shifting from an initial summary mode to a conditional mode (as we found in this study) or whether they show equal flexibility in either direction. Given the discrepant conclusions about change that these modes of assessment sometimes yield, it will be important to probe how experts and novices reconcile these discrepancies in their own judgments. Future research should also examine a wider range of assessment tools, including those that appear to use a summary format (e.g., Conners et al. 1998) and those that also use context-specific items (e.g., Social Skills Rating System/SSRS; Gresham and Elliott 1990).

Although some of our more experienced staff had advanced degrees in clinical psychology, social work, and education, the sample size of this elite group was too small to justify separate analyses. Future studies should explore whether our results extend to more clearly defined "expert" and "novice" groups, such as licensed clinical psychologists and social workers with several years of experience, and individuals with no clinical experience. Our experimental paradigm used clear manipulations of event rates and reaction rates (i.e., increases and decreases of 0.25 and 0.75), partly to parallel past research and partly to ensure that the complex targets showed clear differences in how their reactions changed over time. Although changes of this magnitude have been demonstrated in field studies (see Wright et al. 2011), future studies should determine whether smaller manipulations yield similar findings or perhaps even clearer evidence of experienced raters' greater sensitivity to reactions. Finally, our paradigm used text vignettes that described the targets' social interactions, which may have felt artificial to participants. Future research should examine whether more a naturalistic paradigm that utilizes audio or video stimuli might improve novice and experienced raters' judgments of child behavior.

Overall, our findings suggest that the summary assessments often used in clinical practice and research may not capture context-dependent changes or reveal the effects of clinical experience in this domain. Instead, such measures appear to limit whether raters will report on context-specific behaviors and track complex changes in children's behavioral reactions and their social worlds. Our results also extend previous research that has found experienced staff to be more sensitive than novices to situation-behavior relationships at a single time point (see Dawson et al. 1989). Given our findings that experienced staff are sensitive to reaction patterns when directly asked, it may be possible to develop assessment methods that are more congruent with how these raters naturally perceive behavior. Our results also suggest it may be possible to develop training programs that can improve novice raters' sensitivity to changes in children's reactions to situations. Such research could deepen our understanding of what experienced and inexperienced perceivers are capable of detecting, how summary assessment methods influence their sensitivities, and how to improve assessments of context-specific change.

Acknowledgements This research was supported in part by a National Service Agency AmeriCorps Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the AmeriCorps program. We would like to thank Dennis Calcutt, Noel O'Connor, Maggie Cochran, Carrington Giammittorio, and Pamela Peterson for their assistance. We are especially grateful to Harry Parad, Director of Wediko Children's Services, whose support made this research possible.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington: University of Vermont. doi:10.1007/978-0-387-79948-3_1529.

Block, J. (2010). The five-factor framing of personality and beyond: some ruminations. *Psychological Inquiry, 21*, 2–25. doi:10.1080/10478401003596626.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review, 90*, 105–126. doi:10.1037/0033-295X.90.2.105.

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.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81. doi:10.1016/0010-0285(73)90004-2.

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.

Conners, C. K., Sitarenios, G., Parker, J. D., & Epstein, J. N. (1998). The revised Conners' Parent Rating Scale (CPRS-R): factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology, 26*(4), 257–268. doi:10.1023/A:1022602400621.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. NY: The Free Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Dawson, V. L., Zeitz, C. M., & Wright, J. C. (1989). Expert-novice differences in person perception: evidence of experts' sensitivities to the organization of behavior. *Social Cognition, 7*, 1–30. doi:10.1521/soco.1989.7.1.1.

Dever, B. V., Mays, K. L., Kamphaus, R. W., & Dowdy, E. (2012). The factor structure of the BASC-2 behavioral and emotional screening system teacher form, child/adolescent. *Journal of Psychoeducational Assessment, 30*, 488–495. doi:10.1177/0734282912438869.

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.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273–305. doi:10.1146/annurev.psych.47.1.273.

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.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*, 387–396. doi:10.1037/0033-2909.105.3.387.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association. doi:10.1037/10299-002.

Garb, H. N. (2006). The conjunction effect and clinical judgment. *Journal of Social and Clinical Psychology, 25*, 1048–1056. doi:10.1521/jscp.2006.25.9.1048.

Garcia-Retamero, R., & Dhami, M. K. (2009). Take-the-best in expert-novice decision strategies for residential burglary. *Psychonomic Bulletin & Review, 16*, 163–169. doi:10.3758/PBR.16.1.163.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482. doi:10.1146/annurev-psych-120709-145346.

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., & Elliott, S. N. (1990). *Social skills rating system (SSRS)*. Circle Pines, MN: American Guidance Service.

Grove, W. M. (2005). Clinical versus statistical prediction: the contribution of Paul E. Meehl. *Journal of Clinical Psychology, 61*(10), 1233–1243. doi:10.1002/jclp.20179.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment, 12*, 19–30. doi:10.1037/1040-3590.12.1.19.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software, 15*, 1–11.

Hartley, A. G., Wright, J. C., Zakriski, A. L., & Banducci, A. N. (2013). An experimental analysis of the assessment and perception of behavior change: how summary measures influence sensitivity to change processes. *Psychology, 4*, 1–10. doi:10.4236/psych.2013.41001.

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. doi:10.1016/j.cpr.2008.12.003.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation. In G. Gigerenzer, P. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 209–234). NY: Oxford University Press.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. NY: Routledge.

Hunsinger, M., Isbell, L. M., & Clore, G. L. (2012). 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, 38*, 220–232. doi:10.1177/0146167211424166.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*, 515. doi:10.1037/a0016755.

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.

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. doi:10.1016/j.psychres.2008.08.006.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22. doi:10.1093/biomet/73.1.13.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339–363. doi:10.1037/0033-295X.90.4.339.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reynolds, C. R., & Kamphaus, R. W. (2002). *Behavior assessment system for children—second edition*. Circle Pines: American Guidance Service.

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*(1), 61. doi:10.1037/0022-3514.63.1.61.

Schwarz, N., & Oyserman, D. (2011). Asking questions about behavior: Self reports in evaluation research. In M. Melvin, S. Donaldson, & B. Campbell (Eds.), *Social psychology and evaluation*. New York: Guildford Press. doi:10.1177/109821400102200202.

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: how to decide if someone is an expert or not. *European Journal of Operational Research, 136*(2), 253–263. doi:10.1016/S0377-2217(01)00113-8.

Skinner, E. A., & Zimmer-Gembeck, M. J. (2007). The development of coping. *Annual Review of Psychology, 58*, 119–144. doi:10.1146/annurev.psych.58.110405.085705.

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.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., Rush, J. D., et al. (2009). The meta-analysis of clinical judgment project effects of experience on judgment accuracy. *The Counseling Psychologist, 37*, 350–399. doi:10.1177/0011000006295149.

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.

Trope, Y., Cohen, O., & Maoz, Y. (1988). The perceptual and inferential effects of situational inducements on dispositional attribution. *Journal of Personality and Social Psychology, 55*(2), 165. doi:10.1037/0022-3514.55.2.165.

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.

Westen, D., & Weinberger, J. (2005). In praise of clinical judgment: Meehl's forgotten legacy. *Journal of Clinical Psychology, 61*, 1257–1276. doi:10.1002/jclp.20181.

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., Zakriski, A. L., Hartley, A. G., & Parad, H. W. (2011). Reassessing the assessment of change in at-risk youth: conflict and coherence in overall versus contextual assessments of behavior. *Journal of Psychopathology and Behavioral Assessment, 33*, 215–227. doi:10.1007/s10862-011-9233-x.