THE JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY
ACAMH
*Journal of Child Psychology and Psychiatry* 53:5 (2012), pp 558–574 doi:10.1111/j.1469-7610.2012.02537.x

# Annual Research Review: Embracing not erasing contextual variability in children's behavior – theory and utility in the selection and use of methods and informants in developmental psychopathology

**Melanie A. Dirks,[1] Andres De Los Reyes,[2] Margaret Briggs-Gowan,[3] David Cella,[4] and Lauren S. Wakschlag[4]**

[1]Department of Psychology, McGill University, Montreal, QC, Canada; [2]Department of Psychology, University of Maryland, College Park, MD; [3]School of Medicine, University of Connecticut, Farmington, CT; [4]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

This paper examines the selection and use of multiple methods and informants for the assessment of disruptive behavior syndromes and attention deficit/hyperactivity disorder, providing a critical discussion of (a) the bidirectional linkages between theoretical models of childhood psychopathology and current assessment techniques; and (b) current knowledge concerning the utility of different methods and informants for key clinical goals. There is growing recognition that children's behavior varies meaningfully across situations, and evidence indicates that these differences, in combination with informants' unique perspectives, are at least partly responsible for inter-rater discrepancies in reports of symptomatology. Such data suggest that we should embrace this contextual variability as clinically meaningful information, moving away from models of psychopathology as generalized traits that manifest uniformly across situations and settings, and toward theoretical conceptualizations that explicitly incorporate contextual features, such as considering clinical syndromes identified by different informants to be discrete phenomena. We highlight different approaches to measurement that embrace contextual variability in children's behavior and describe how the use of such tools and techniques may yield significant gains clinically (e.g., for treatment planning and monitoring). The continued development of a variety of feasible, contextually sensitive methods for assessing children's behavior will allow us to determine further the validity of incorporating contextual features into models of developmental psychopathology and nosological frameworks. **Keywords:** Methodology, assessment, development, ADD/ADHD, disruptive behavior, situation specificity, informant discrepancies.

## Introduction

Given there is no biological or behavioral marker that definitively indicates the presence of clinically impairing psychological syndromes in children or adolescents (De Los Reyes, 2011; Kraemer et al., 2003), the collection of data from multiple sources is, by necessity, the gold standard for measuring developmental psychopathology (Hunsley & Mash, 2007). Thus, clinicians and researchers are tasked with two key assessment decisions: (a) How, and from whom, should information be collected? (b) How should the resulting data be integrated? Little consensus exists on how to make these important choices. The Diagnostic and Statistical Manual of Mental Disorders-IV-TR (DSM-IV, American Psychiatric Association, 2000), for example, provides minimal guidance concerning what information should be obtained to guide diagnostic decision-making (Hudziak, Achenbach, Althoff, & Pine, 2007).

In this paper, we critically review evidence for the selection and use of multiple methods and informants to assess psychopathology in children and adolescents (herein referred to as 'children'), focusing on the DSM disruptive behavior syndromes (oppositional defiant disorder, ODD, and conduct disorder, CD) and attention deficit/hyperactivity disorder (ADHD), because the empirical knowledge base concerning multi-method and multi-informant measurement is most substantial for these syndromes. Our goal is not to provide a comprehensive cataloguing of tools and techniques for assessing these syndromes, as others have done this work (e.g., Hunsley & Mash, 2008). Rather, our objective is to examine: (a) the bidirectional linkages between conceptual models of childhood psychopathology and commonly used assessment techniques; and (b) current knowledge concerning the utility of different methods and informants for making diagnoses and planning and monitoring treatment across developmental stages.

Conflict of interest statement: Dr. Briggs-Gowan receives royalties from Pearson Assessment, the publisher of the Infant-Toddler Social Emotional Assessment (ITSEA).

Assessment practices – which include not only the instruments used but techniques for quantifying, summarizing, analyzing, and interpreting the resulting information – should maximally fit theoretical models of the clinical phenomena under consideration (McFall, 2005). Psychological syndromes are theoretical models advanced to explain patterns in children's functioning (Kendell & Jablensky, 2003), and these conceptualizations are tested by measuring referents, which are the observable phenomena that are the manifestations of the underlying construct (McFall & Townsend, 1998). As such, selecting measures and analytic approaches stakes a theoretical claim. For example, integrating data across different informants' reports to form one score reflects an implicit conceptualization of the underlying syndrome as a unitary construct of interest that generalizes across settings (Gomez, Burns, Walsh, & de Moura, 2003). Thus, beyond psychometric considerations, choosing among measures is inherently a theory-based process that necessitates thoughtful evaluation of the nature of the phenomena under consideration.

For this reason, there must be an ongoing dialogue between assessment and theory, such that theoretical and nosological frameworks are continuously refined to accommodate the data resulting from developments in assessment techniques. The current psychiatric nosology has advanced our understanding of childhood psychopathology in many important ways (Angold & Costello, 2009). Its problems, however, have also been documented. Critically, the existing diagnostic categories do not provide adequately defined phenotypes for studies of genetic contributions to psychiatric symptomatology (e.g., Ginsburg et al., 1996), nor are they easily integrated with findings from clinical neuroscience (Insel et al., 2010). The DSM taxonomy yields groupings that are merely descriptive, highly heterogeneous, and markedly overlapping. As researchers seek to identify increasingly specific causal mechanisms, it has become apparent that alternative approaches to organizing behavioral and emotional dysfunction are needed (Sanislow et al., 2010).

One promising possibility is to examine functional characterizations reflecting the different circumstances in which children's symptoms manifest (De Los Reyes, Henry, Tolan, & Wakschlag, 2009; Wright & Zakriski, 2001). Mounting evidence suggests that children's behavior varies reliably and meaningfully across interpersonal situations (Dirks, Treat, & Weersing, 2007a; Wright, Zakriski, & Drinkwater, 1999). These behavioral differences, in combination with informants' unique perspectives on children's behavior, are at least partly responsible for the discrepancies that occur when different raters are asked to report on children's behavior and psychological symptoms (De Los Reyes & Kazdin, 2005; Dumenci, Achenbach, & Windle, 2011). Variability in assessments of children's behavior – across both specific interpersonal situations and more broadly construed settings (e.g., home, school, clinic), and as judged by different individuals – has often been considered something to be erased, in order to identify the 'true' dispositions underlying actions (see Wright, Zakriski, Hartley, & Parad, 2011). We, on the other hand, take the perspective that such differences should be embraced, as they will contribute to our understanding of psychopathology. There may be both conceptual and practical benefits to revising our theoretical models to incorporate commonly observed contextual variations in children's behavior generally and symptom presentations in particular. We suggest that symptoms occurring in different situations or settings, or as perceived by different informants, may constitute distinct phenotypes, and that our ability to understand those phenotypes holds the promise of advancing the diagnosis and treatment of psychopathology.

Advancing understanding of why behavioral variability occurs and what it can reveal about the heterogeneity of behavioral disorders in children will help both clinicians and researchers to collect the most relevant information and to use those data efficiently. These issues of utility, or the extent to which an assessment practice contributes to improved clinical decision-making (Hunsley & Mash, 2007), must also inform assessment choices (McFall, 2005). During an assessment, each instrument, informant, and data-analytic strategy should show evidence of incremental validity, contributing uniquely to the goal(s) of the process (Hunsley & Meyer, 2003). The rule of parsimony should prevail unless there is empirical evidence that 'more is better,' yet too often more intensive practices are adopted in the absence of compelling evidence for their added value (Cella, Gershon, Lai, & Choi, 2007; Dirks & Boyle, 2010). This is a critical challenge for the field, as incorporating more complex procedures without demonstrated utility increases burden on families and may exacerbate clinicians' resistance to the incorporation of standardized measurements into clinical practice (Johnston & Murray, 2003).

## Choosing among methods in the assessment of child psychopathology

In this section, we review how these considerations of theory and utility do, and should, guide clinicians and researchers as they choose among three of the major approaches to assessing psychopathology: rating scales, interviews, and observational procedures. These three strategies share a key limitation, which is that they only provide access to information that can be reported or seen. There are important processes in the etiology and maintenance of ADHD, ODD, and CD that can only be assessed through biological or performance-based tasks (e.g., cognitive functioning in ADHD; Pelham, Fabiano, & Massetti,

2005). Although such tools have yielded significant insights into these syndromes, their clinical utility has yet to be widely established. This situation is likely to change. For example, recent work incorporates a performance-based test of interpretation biases into a treatment protocol for adolescent mood disorder (Lothmann, Holmes, Chan, & Lau, 2011), and as the Research Domain Criteria initiative advances understanding of the underlying etiological mechanisms of psychiatric disorder (Insel et al., 2010), performance-based and biological assessment approaches will likely become more common. Yet even as the validity of such tasks for clinical purposes becomes increasingly established, there will always be need for valid and reliable indices of children's observable functioning. Understanding of the cognitive, biological, and social mechanisms contributing to psychopathology is advancing rapidly, but the complexity of both the pathways leading to children's behavioral and emotional dysfunction and the resulting phenotypes makes it unlikely that the field will reach a point when valid and reliable reports of phenomenology will play no role in diagnosis (Kendler, 2005). Moreover, they will always be important for treatment planning and monitoring (Pelham et al., 2005). Given their current and future importance for the assessment of psychopathology, it is essential that we continually evaluate the theoretical underpinnings and clinical utility of rating scales, interviews, and observational procedures.

### *Underlying theoretical models of rating scales, interviews, and observational procedures*

Rating scales and interviews, which can be unstructured, respondent based (i.e., structured) or interviewer based (i.e., semi-structured), are the most widely used tools for assessing childhood psychopathology (e.g., Hunsley & Mash, 2008). Naturalistic and laboratory observational procedures are also used in the assessment of ADHD, ODD, and CD (Frick & McMahon, 2008; Pelham et al., 2005), often to corroborate evidence provided by rating scales or interviews (McConaughy et al., 2010). Diagnostic observation procedures, however, are specifically designed to generate unique information to be incorporated into clinical decision-making, by engaging families in standardized laboratory procedures that 'press' for the range of clinically salient behaviors (Lord et al., 2000).

Although the methodologies differ, rating scales, interviews, and observational procedures, as they are typically used, reflect similar underlying theoretical models of child psychopathology. Specifically, they emphasize psychopathology as a trait that will generalize across situations (see Wright et al., 2011). In general, rating scales ask informants to make global judgments about the frequency or intensity of symptoms (Dirks et al., 2007a; McDermott, 1993; Wright et al., 1999). For example, the widely used Child Behavior Checklist (Achenbach & Rescorla, 2001) asks parents to evaluate the extent to which statements such as 'argues a lot,' 'talks too much,' and 'threatens people' are true of their child. Such ratings emphasize overall frequency or require a global trait judgment without explicit reference to the interpersonal circumstances in which symptoms are occurring. Diagnostic interviews also tend not to solicit information about context, except in cases where the DSM-IV criteria explicitly reference contextual antecedents. For example, to be diagnosed with ADHD, impairment must be present in two settings, and as such, diagnostic interviews often query whether symptoms or impairment occur at home, at school, or in other contexts (e.g., the Child and Adolescent Psychiatric Assessment, CAPA, Angold & Costello, 2000).

In contrast to interviews and rating scales, observational measures provide a significant amount of contextual information, both at the setting and situation level. Observations can take place at home, school, or in the clinic, three discrete settings that present different demands. Moreover, it is possible to observe the specific interpersonal circumstances that precede behaviors. Often, however, this information is disregarded when these procedures are used to obtain decontextualized frequency counts of behavior (Wright et al., 2011). In addition, behavior may only be assessed in one setting, under the assumption that it will generalize to others, which may not be the case (see Gardner, 2000). Thus, as they are typically used, observational approaches are consonant with a theoretical model similar to that described for rating scales and interviews. Behaviors are the referents of the underlying psychopathology, without regard to the interpersonal situations in which they are embedded. Situational information is available, but not considered.

*The role of interpersonal context in developmental psychopathology.* Increasingly, however, there is emphasis on the importance of context for diagnostic assessment (Drabick, 2009). The situations in which symptoms are elicited can provide important clues concerning the presence and severity of the syndrome. Pervasiveness, the extent to which symptoms are displayed across situations, has been identified as a key indicator of psychopathology in children (Angold & Costello, 2000), an idea that has been incorporated into some measures. For example, the CAPA uses the number of activities in which symptoms occur as a marker of intensity (Angold & Costello, 2000). The Adjustment Scales for Children and Adolescents (McDermott, 1993) operationalizes psychopathology as the occurrence of symptoms across discrete situations. Teachers are asked to identify how a child responds in a particular interpersonal circumstance (e.g., when given correction) from a menu of behaviors (e.g., 'takes correction without fuss,' 'takes correction badly, [such as]

sulky muttering, expressions, etc.,') and the presence of a clinically concerning syndrome is determined based on the number of situations in which the corresponding behavior occurs. Clinically significant oppositionality, for example, is identified when a child is reported to engage in the related behaviors in six or more situations (also see Dodge, McClaskey, & Feldman, 1985; DuPaul & Barkley, 1992).

A second way in which situation may modify the clinical significance of a behavior is through the principle of developmental expectability (Wakschlag, Tolan, & Leventhal, 2010). Some behaviors are more likely, or expectable, in particular interpersonal contexts (Cole, Martin, & Dennis, 2004) and behavior that occurs in an expected situation (e.g., a preschooler displaying aggression during a toy dispute) may not be as clinically concerning as behavior that occurs in atypical circumstances (e.g., aggression by a preschooler that appears to come 'out of the blue;' Wakschlag et al., 2010). In this way, the referent of the underlying pathology is the behavior tied to its interpersonal and/or broader contextual antecedents (e.g., setting). This idea has been incorporated into some measures for some symptom types, with the most common example being the discounting of aggression toward siblings as a symptom of ODD or CD.

Both pervasiveness and expectability involve considering contextual information at the level of the referent. Alternatively, the underlying syndrome could be conceptualized to incorporate contextual variability in the manifestation of symptoms (Wright & Zakriski, 2001). Within groups of children experiencing each of ADHD, ODD, and CD, there is variability in the number and types of situations in which they exhibit symptoms (e.g., DuPaul & Barkley, 1992; Matthys, Maassen, Cuperus, & van Engeland, 2001). Although little work has examined this issue, it is possible that stable patterns of situation-symptom contingencies may underlie, or cut across, the existing diagnostic categories. For example, Wright and Zakriski (2001) found that within a group of boys exhibiting clinically significant conduct problems, two distinct subgroups emerged, differentiated by situational variability in aggressive behavior: One group was perceived by teachers to show elevated aggression only in response to aversive events with peers, whereas the second was perceived to engage in elevated aggression in response to all interpersonal situations. Within this theoretical framework, then, not only will symptomatology show variability across situations at the level of the child, but there will be stable patterns of situations and responses that differentiate groups of children. For example, a child who displays oppositionality only with a parent would be considered different from a child who displayed such behaviors only with peers. This approach is evident in the distinction in developmental psychopathology research between reactive and proactive aggression. These two behaviors differ in the eliciting situations: Reactive aggression occurs in the context of perceived provocation, frustration, or threat. In contrast, proactive aggression is planned behavior intended to help achieve a desired outcome (Crick & Dodge, 1996). These types of aggression have been associated with unique developmental pathways (e.g., Brendgen, Vitaro, Boivin, Dionne, & Perusse, 2006), suggesting that the incorporation of contextual information may contribute to the identification of more precise etiological mechanisms.

To be consistent with a theoretical model incorporating contextual patterning, the referents must tie symptoms to context, and this information should be maintained when data is aggregated or summarized. Some rating scales and interviews contain contextualized items (e.g., 'argues when denied own way,' Wright et al., 2011), but when these items are added together to form a total score this information is lost. The Behavior-Environment Transactional Analysis (Wright & Zakriski, 2001) is an example of an instrument that maintains situational patterning of behavior by asking informants to report how often children encounter specific social events and how they respond in these circumstances, and then capturing these situation-behavior contingencies in the scoring. For example, children receive separate scores for aggressive behavior in response to aversive events occurring with adults and aversive events occurring with peers. Similarly, there are a number of inventories that assess children's management of key interpersonal situations, such as responding to conflict with a friend (Rose & Asher, 1999), and peer provocation (Dirks, Treat, & Weersing, 2007b). To date, such measures have not been widely integrated into clinical research or practice. The data they yield, however, could prove valuable for these purposes by providing detailed information about the specific circumstances under which behavioral dysfunction is occurring.

### *Issues of utility in the use of rating scales, interviews, and observational procedures*

Ultimately, bringing our measurement approaches in line with data concerning the situational specificity of youth symptomatology should pay dividends for clinical decision-making. When evaluating the utility of an assessment practice, it is important to consider who is being assessed and why, as different methods will be more or less informative depending upon developmental stage (Silverman & Ollendick, 2005), and will be better suited to some purposes than others (Angold & Costello, 2009). We next consider the utility of rating scales, interviews, and observations for different clinical tasks, highlighting circumstances under which the incorporation of contextual information may be particularly useful.

*Diagnostic decision-making.* Many clinicians rely on unstructured interviews for diagnostic purposes (Jensen-Doss & Hawley, 2010). A recent meta-analysis indicates that case classifications based on these evaluations show limited agreement with those yielded by standardized interviews (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009), a difference that may be partly due to the demands of clinical practice. Jensen and Weisz (2002) compared clinician-generated diagnoses to those obtained through a structured interview and found that the latter were more likely to result in no diagnosis, which may reflect the reality that clinicians have to assign a diagnosis to have services authorized. On average, standardized interviews also generated more diagnoses for a given child, a discrepancy that may relate to time limitations that understandably force clinicians to focus on the primary concern. Although these differences make sense given the constraints of clinical practice, there is evidence that structured interviews are more comprehensive and reliable than unstructured interviews (see Garb, 2007; for review). Only a few studies have examined whether greater structure affords increased validity (Garb, 2007); available evidence, however, suggests that standardized interviews yield more valid classifications than unstructured interviews (see Jensen-Doss & Hawley, 2010; for review). This difference may be due, in part, to the minimization of biases affecting the unstructured collection of diagnostic information, (e.g., selectively obtaining information that confirms initial impressions; Garb, 2007), and other influences on clinicians' judgment (e.g., therapeutic orientation, Pottick, Kirk, Hsieh, & Tian, 2007). Such data clearly indicate the benefits of incorporating standardized assessments into clinical practice.

A primary reason that clinicians are reluctant to use standardized tools is that they view them as impractical (Jensen-Doss & Hawley, 2010). In this regard, interviews are considerably more onerous than rating scales as they are lengthier, and, in the case of semi-structured interviews, must be administered by a trained individual. It is critical, then, that the scientific benefits of interviews, compared to rating scales, compensate for this additional burden. The relative utility of interviews will vary as a function of the purpose of the assessment. For example, for researchers wishing to quantify symptomatology, available evidence suggests that briefer rating scales perform as well as respondent-based interviews in community samples (Dirks & Boyle, 2010). Similarly, their equivalence has been demonstrated when diagnoses are being made to estimate prevalence rates in the general population (e.g., Boyle et al., 1997). Under these conditions, as long as false positives and false negatives are roughly balanced, inferences will not be affected (Costello, Egger, & Angold, 2005).

In the clinic, however, diagnoses are tied to treatment, making accurate identification of cases critical. A number of studies have demonstrated that rating scales perform as well as structured interviews for diagnosing ADHD (see Johnston & Mah, 2008; Pelham et al., 2005). However, further work is needed to assess the generalizability of these findings, particularly among samples of youth seeking clinical services. There appears to be less work comparing interview- and rating scale-based diagnoses of ODD and CD, but two investigations have suggested that rating scales perform comparably to structured (Edelbrock & Costello, 1988) and semi-structured interviews (Grayson & Carlson, 1991).

This research provides preliminary evidence for the possibility that ADHD, ODD, and CD could be accurately diagnosed with briefer assessments. Further support for the potential of using shorter measures comes from data indicating that, although the DSM-IV weights all symptoms of these syndromes equally, some symptoms are more predictive of a diagnosis than others (e.g., Frick et al., 1994; Gelhorn et al., 2009; Power, Costigan, Leff, Eiraldi, & Landau, 2001). Such findings suggest the possibility of paring down assessment items. Building on this idea, psychometric advances, particularly item-response theory (IRT), have facilitated the development of computerized adaptive testing (CAT), an individualized approach to measurement that greatly reduces the number of questions needed to assess accurately the construct of interest (Cella et al., 2007). Applications of CAT to the assessment of psychopathology have begun recently (Reise & Waller, 2009) and wider use of this technique will contribute to the development of more efficient assessment batteries. This dissemination will also advance basic knowledge of developmental psychopathology, as this approach could provide information about which symptoms, as rated by which informants, are most predictive of clinically significant syndromes.

Although there is evidence suggesting that briefer assessments may yield comparable classification, it is also important to consider other types of information essential for diagnostic decisions, as well as whether a trained interviewer may be necessary to gather these data. For example, an interviewer may be able to obtain more precise estimates of the onset and duration of symptoms, which may prove important given evidence for the different trajectories associated with early- versus late-onset CD (Frick & McMahon, 2008). Empirical tests of the incremental validity of interviews should consider all of the information being gathered, to pinpoint more precisely the conditions under which interviews yield significant added value.

This work must also consider developmental level, as it is likely that the need for more intensive assessments, such as interviews or diagnostic observation, will vary across childhood. During some developmental periods, it may be difficult for someone without specialized training to determine

whether a behavior is clinically concerning. For example, aggression and oppositionality commonly occur in preschoolers. Thus, the presence of these behaviors per se may not be clinically informative as it is in older children, making reliance on reports of behavioral frequency inadequate during this developmental period (Wakschlag et al., 2010). To address the challenges associated with disentangling clinically significant disruptive behavior from normative misbehavior at this age, Wakschlag et al. (2008a,b) developed the Disruptive Behavior Diagnostic Observation System (DB-DOS). This standardized diagnostic observation moves beyond simple behavioral counts by using ordinal ratings to code clinical concern. These judgments are based on the quality of behavior, its age appropriateness, and, importantly, the context in which it is occurring. For example, saying 'no' in response to a request to clean up is developmentally expectable, but a 'reflexive' no across a range of circumstances is not and is thus coded as clinically concerning (Wakschlag et al., 2007). Early evidence from the DB-DOS suggests that examining the expectability and pervasiveness of preschoolers' disruptive behavior may have incremental clinical utility, suggesting the potential value of systematically incorporating contextual information for accurate identification of clinical syndromes (Wakschlag et al., 2008a).

*Treatment planning and monitoring.* Incorporation of contextual features could also be helpful for planning or monitoring progress in treatment. To date, little research has examined treatment utility, the extent to which an assessment contributes to beneficial intervention outcomes (Mash & Hunsley, 2005). Some studies, however, have shown that functional-analytic assessments, which focus on understanding the conditioning of symptom expression, are associated with greater improvement in treatment (Haynes, Leisen, & Blaine, 1997), suggesting the possible utility of incorporating situation-level variability into measures and maintaining it in scoring algorithms. Similarly, assessing change in overall rates of behavior may obscure important differences in children's behavior in specific social situations. Wright et al. (2011) found that over the course of a therapeutic summer-camp program, children's average level of prosocial behavior increased and mean aggression decreased. Closer inspection revealed, however, that children's aggressive behavior actually increased significantly in response to provocation by a peer, and prosocial behavior in this situation decreased. These findings suggest that although there are a number of rating scales sensitive to change in treatment for ADHD, ODD, and CD (Frick & McMahon, 2008; Johnston & Mah, 2008), such global assessments may provide an incomplete accounting of behavioral change. Children's behavior may improve in some situations, but worsen or show no change in others.

Although much more research is needed to determine the generalizability of these findings, particularly within the context of widely implemented interventions for childhood disorders, preliminary evidence points to the value of situation-specific measurement in the context of intervention planning and delivery. Observational methods would seem ideally suited to this task, and work is ongoing to increase the feasibility of these approaches for clinical use (Pelham et al., 2005; Wakschlag et al., 2008b). It is also possible to translate the knowledge gained from more intensive methodologies into briefer instruments. The nuanced information about situation-behavior patterning gleaned from qualitative interviewing or observational paradigms can provide the foundation for the construction of contextually and developmentally sensitive rating scales (e.g., Dirks, Treat, & Weersing, 2011b; Wright & Zakriski, 2001). For example, Wakschlag, Briggs-Gowan and colleagues have 'translated' constructs about behavioral qualities salient to identification of disruptive behavior at preschool age from direct observation during the DB-DOS to a paper and pencil measure. This Multidimensional Assessment of Preschool Disruptive Behavior queries multiple facets and contexts of behavior in order to distinguish normative from clinically concerning occurrence (Wakschlag et al., 2011). In general, such measures maintain valuable contextual information, but may be more broadly useful given how much easier they are to administer.

The use of contextualized measures of psychopathology holds considerable promise for clinical practice, and we advocate here for research that examines empirically the clinical utility of such assessments. To begin, incorporating existing situation-based inventories of children's behavior (e.g., Dirks et al., 2007b; Rose & Asher, 1999) into intervention studies will provide preliminary evidence concerning whether the inclusion of contextual information provides a more nuanced view of behavior change. Next, a more systematic approach could be taken to assess the treatment utility of these measures. In general, randomized studies of measurement approaches have been rare; however, such investigations would provide significant information about the relative utility of different assessment strategies. For example, using this approach to assess the incremental validity of systematically tiered levels of methods would provide an empirical knowledge base for decision-making about inclusion of various levels of measurement. Although this type of work would be labor intensive, such efforts would be justified by the strength of the resulting inferences and implications for treatment, and the proposed studies offer a promising opportunity for researchers and clinicians to collaborate in ways that would enhance assessment, theory, and intervention.

## Use of multiple informants in the assessment of developmental psychopathology

As the field works to develop rating scales and interviews that focus explicitly on the contextual patterning of children's behavior, clinicians and researchers who use these methods will continue to face a second choice point: Who should complete them? It is generally recommended that data be collected from more than one informant (Hunsley & Mash, 2007) and typical raters include the children themselves, their parents, and their teachers. It has been widely documented that the agreement between any two of these individuals will be low to moderate (Achenbach, McConaughy, & Howell, 1987; De Los Reyes & Kazdin, 2005). To make sound choices about which informants to ask, and ultimately, to make sense of the resulting data, it is essential to understand why these discrepancies arise.

### *Reasons for informant discrepancies*

Historically, random error, which can result from a number of different factors (e.g., differing interpretations of the anchors on a rating scale), has been viewed as the principal reason informants diverge (De Los Reyes, 2011). Several lines of evidence, however, suggest that this is not the case. First, different informants provide reports of children's behavioral problems that are reliable and valid (De Los Reyes, 2011). Second, reports by different raters often share unique associations with a number of indices of youth functioning, both concurrently and longitudinally (see Burt, McGue, Krueger, & Iacono, 2005; Collishaw, Goodman, Ford, Rabe-Hesketh, & Pickles, 2009) and some research suggests that the variance unique to informants may share stronger associations with criterion variables than the variance shared between them (Dirks, Boyle, & Georgiades, 2011a; but see Van Dulmen & Egeland, 2011). Third, discrepancies between informants are stable over time (e.g., De Los Reyes, Alfano, & Beidel, 2010) and show high levels of internal consistency (De Los Reyes et al., 2011a).

*Reason 1: Informants' unique perspectives.* Given such findings, it is likely that systematic differences between raters are playing a bigger role in informant disagreement. Some of these are sources of error: Factors that cause raters to consistently report particular symptoms unconfirmed by other sources. A significant amount of work has focused on detailing such biases (see De Los Reyes & Kazdin, 2005), including contrast effects, such that the behavior of one sibling influences perceptions of the other (Simonoff et al., 1998); and halo effects, in which estimates of a given behavioral problem (e.g., ADHD), are inflated in the presence of other symptom types (e.g., ODD; Abikoff, Courtney, Pelham, & Koplewicz, 1993).

Some of the differences between informants' perceptions, however, likely reflect variability in the meaning or interpretation of a particular behavior or symptom across contexts (De Los Reyes et al., 2009; Dirks, Treat, & Weersing, 2010). Research has shown that thresholds for the acceptability of children's behavior vary as a function of cultural factors (see Weisz, McCarty, Eastman, Chaiyasit, & Sunwanlert, 1997), and at a more micro level, these thresholds likely also vary across settings. In school, for instance, teachers must handle the demands of managing a classroom, and under these circumstances, behaviors that are often considered assertive, such as questioning rules and perceived unfair treatment (Gresham & Elliott, 2008), may be construed as oppositional. As such, informant discrepancies may be capturing, in part, differences in the types of behaviors that are problematic in a given context from the perspective of a particular informant (Dumenci et al., 2011). When considered within this framework, variability in informants' ratings is not a problem, but an opportunity to learn about children's adaptation in various settings. Disentangling the extent to which informant discrepancies reflect factors resulting from rater characteristics and genuine differences in the meaning of a behavior across settings will be an important focus for future research.

*Reason 2: Situation specificity of children's behavior.* In addition to informant characteristics and perspective, researchers have hypothesized that the marked situation specificity of children's behavior is a key contributor to inter-rater discrepancies (Achenbach et al., 1987; De Los Reyes & Kazdin, 2005). Previously, support for this supposition has been limited to the indirect evidence that there is greater agreement between informants in the same setting (e.g., peers and teachers) than informants in different settings (e.g., parent and teacher; Achenbach et al., 1987). Two recent studies, however, provide more direct corroboration. De Los Reyes et al. (2009) used the DB-DOS to examine the associations between preschoolers' disruptive behavior observed in two interpersonal contexts – interacting with an examiner and interacting with a parent – and different informant ratings. Results indicated that observed disruptive behavior with the parent was associated with parent, but not teacher, ratings of disruptive behavior, whereas observed disruptive behavior with the examiner was associated with teacher, but not parent, ratings of disruptive behavior, a pattern that indicates that contextual variability in children's behavior is 'real,' and not merely an artifact of rater characteristics. In a second study, Hartley, Zakriski, and Wright (2011) found that greater similarity in the types of interpersonal events children experience at home and school predicted increased agreement between parent and teacher reports of their aggressive behavior,

suggesting that some of the discrepancy between parent and teacher reports might be attributable to differences in the social situations children encounter in each context. Greater situational similarity likely leads to increased consistency of behavior across contexts, which then contributes to greater convergence between informants. In both of these studies, then, variability in children's behavior across interpersonal contexts, defined by both interaction partner and interaction type, contributes substantially to inter-rater agreement.

### *Theoretical implications of informant discrepancies*

Such evidence that the differences between informants reflect meaningful variation is inconsistent with the historical emphasis in developmental psychopathology on the agreement between raters (see Hartley et al., 2011). This expectation of convergence is consistent with a theoretical model of psychopathology as a trait that generalizes across contexts (see Rowe & Kandel, 1997): a given syndrome should manifest in the same way across settings and situations, and be perceived in the same way by raters. Within this conceptual framework, each informant is thought to provide an alternate sample of the indicators of the underlying construct. As noted by McFall and Townsend (1998) 'if the construct is a good one, these different sampling methods should yield convergent evidence' (p. 317). If measurement of the referents is adequate, and if the referents reflect the same underlying construct, informants' ratings should converge.

The presence of significant discrepancies between raters, then, signals one of two issues. Given the characteristics and perspective unique to each informant, inter-rater discrepancies may reflect differences in understanding of the referents. There may be variability between parents and teachers, for example, in their judgments of the nature and severity of behaviors that would warrant ratings of 'often forgetful,' 'often leaves seat,' and 'easily distracted.' In this case, it may be possible to reduce inter-rater discrepancies by providing tighter specification of symptoms. If, however, differences in ratings are at least partly driven by informants' access to different behavioral samples, as well as differences in the meaning of a behavior in a given context, then the issue may lie with the overarching theoretical construction. Rather than reflecting a unitary syndrome, it may be that reports by different informants are representative of different underlying constructs; for instance, teacher-reported ODD may be a different construct than parent-reported ODD (see Drabick, Bubier, Chen, Price, & Lanza, 2011; Drabick, Gadow, & Loney, 2007), a conceptualization that maps on to findings, reviewed earlier, that there may be functional differences between children in the manifestation of psychopathology (e.g., Wright & Zakriski, 2001). For example, children who have behavioral difficulties only in interactions with peers may be identified by teachers, but not parents (De Los Reyes et al., 2009).

*Implications of theoretical models of informant discrepancies for data aggregation strategies.* Clarifying the underlying theoretical model of informant discrepancies is critical because it informs the selection of strategies used to combine multi-informant data. Many of the strategies used to aggregate data are inconsistent with the burgeoning evidence that source-specific variability is meaningful. The 'or' rule counts a symptom (or diagnosis) as present if it is endorsed by any informant, making no distinction between children for whom there is agreement about symptoms or classification and those for whom there is disagreement (Dirks et al., 2011a). Adding symptoms identified by different informants together also does not distinguish between raters; a child who had two symptoms reported by a parent and six by a teacher would be treated the same as one who had eight symptoms reported by a parent and none by a teacher (Holmbeck, Li, Schurman, Friedman, & Coakley, 2002). Alternatively, the 'and' rule emphasizes convergence of information; symptoms (or diagnoses) 'count' only when informants agree. Similarly, latent constructs that combine data provided by multiple raters reflect the variance shared between informants, with unique information relegated to the error terms (Holmbeck et al., 2002), although it is possible, with careful selection of informants, to model inter-rater discrepancies meaningfully (Kraemer et al., 2003).

Treating raters as equivalent, or discarding the differences between them as error, will result in the loss of valuable information about children's current impairment and ultimate prognosis, leading a number of authors to suggest that information provided by raters should be maintained separately (e.g., Drabick et al., 2007; Offord et al., 1996). This source-specific approach is consistent with a theoretical model that indicates that variability across informants' ratings is consequential. It also assumes, however, that the agreement between raters is not informative (Baillargeon, Boulerice, & Tremblay, 2001), a problematic premise for at least two reasons. First, the variability shared among informants is consistently associated with outcomes of interest (e.g., Cole, Martin, Powers, & Truglio, 1996; Perren, Von Wyl, Stadelmann, Burgin, & Von Klitzing, 2006) suggesting that it is not occurring purely by chance. Second, there may be differences between children identified as exhibiting a clinical syndrome by multiple informants compared to one informant, variability that will not be apparent if ratings are kept separate (e.g., Ho et al., 1996).

What is needed, then, are strategies that capture both the convergence and divergence among raters. One approach is to differentiate between children

identified as having a clinically impairing syndrome by one or multiple informants. The ADHD and Disruptive Behaviors Workgroup has suggested this approach for DSM-5, recommending the use of a severity index of ODD based on the pervasiveness of symptoms across contexts (Drabick, 2011). Because informant typically serves as a proxy for setting (Drabick, 2011), in practice, this approach, which is similar to the DSM-IV specification that impairment must be present in two settings for a child to be diagnosed with ADHD, would often mean children identified by only their parent (or a teacher) would be seen as having a less severe presentation than those identified by both.

Research, however, does not unequivocally support this framework for either ODD or ADHD. For example, Drabick et al. (2007) compared boys in three groups: those who met criteria for ODD based on maternal report only, teacher report only, or report by both informants (combined). To provide support for the hypothesis that the combined group was the most severely impaired, their functioning would have to be significantly poorer than both of the single-informant groups. This pattern emerged for two variables out of eighteen. In a second study, Munkvold, Lundervold, Lie, and Manger (2009) found that a combined group was significantly more impaired, as rated by both parents and teachers, than parent- and teacher-only groups, and had more CD symptoms, as rated by teachers. The combined group was not the most severe on seven other variables, however, and was identified using the 'and' rule for symptoms, which resulted in the identification of a relatively small (.2% of 7007 children), and thus possibly unrepresentative, group.

Evidence for the hypothesis that 'pervasive' ADHD identified by both parents and teachers represents a more severe presentation than 'situational' ADHD identified by only one of these informants is also not clear cut (see Costello, Loeber, & Stouthamer-Loeber, 1991; Ho et al., 1996). Some work has shown that pervasive ADHD is associated with poorer functioning than situational ADHD on a number of objectively measured indices, including inhibitory control and response reengagement (Schachar, Tannock, & Logan, 1993), IQ (Schachar, Rutter, & Smith, 1981), and objectively measured levels of hyperactivity (Tripp & Luk, 1997). Other investigations, however, are not consistent with this pattern, finding no significant difference across groups on the latter two variables (e.g., Costello et al., 1991; Rapoport, Donnelly, Zametkin, & Carrougher, 1986; Rettew et al., 2011).

Two recent investigations have relevance for the applicability of this conceptualization to CD, comparing children identified as having clinically concerning conduct problems by parents only, teachers only, or both parents and teachers (combined). The first found that the combined group had significantly lower IQ scores and significantly greater parent-rated impairment than the other two groups (Rettew et al., 2011). The magnitude of the difference between the combined and parent-only groups on the impairment rating was small, however, and teacher-ratings of impairment did not differ between the combined and teacher-only groups. The second found no difference between the groups on a number of longitudinal outcomes, including criminality, substance use, anxiety, and depression, although the small number of children per group may have limited analytic power (Fergusson, Boden, & Horwood, 2009).

Taken together, available data are not clearly consonant with a model positing cross-setting pervasiveness as a marker of syndrome severity. There are not enough studies assessing the patterning of correlates and outcomes across syndromes identified by different informants to draw firm conclusions about ODD and CD, and although more data are available concerning ADHD, interpretations are complicated by small sample sizes, underrepresentation of girls, and differences across studies in the definition of situational hyperactivity (i.e., are children identified by parents only or teachers only considered separately or collapsed into one group). As we await further research to elucidate this issue, two themes emerge from the extant literature. First, clinically significant syndromes identified by only one of parents or teachers are associated with substantial impairment and should not be discounted (see Drabick, 2011; Fergusson et al., 2009). Clinicians may wish to investigate carefully whether ADHD reported by parent only would be better characterized as a disruptive behavior problem, given data suggesting that these children (a) are not distinguishable from children with antisocial behavior problems on a number of indices, including family relationships and IQ (Ho et al., 1996), (b) do not demonstrate the same deficits in executive control exhibited by children with ADHD identified by a teacher (Schachar et al., 1993), and (c) may have better long-term prognoses (Mannuzza, Klein, & Moulton, 2002). Second, the assumption made by a cross-setting severity index is that what matters is the number of settings in which children are impaired, but previous work suggests that *which* settings is also critical information, as children with clinically significant syndromes identified by parents appear different from those identified by teachers as well as from those identified by both.

### *Utility and the use of multi-informant data: choosing among aggregation strategies and informants*

*Aggregation strategies.* If there are important differences among groups of children identified via different combinations of informants, then there should be utility in maintaining this patterning

during clinical decision-making. Researchers have adopted a number of different approaches to capture this information analytically. Laird and Weems (2011) suggested constructing regression models that assess whether the interaction between informants explains additional variance, after accounting for the prediction afforded by the separate ratings. Kraemer et al. (2003) advocated using a principal components analysis to parse explicitly the variance between informants into three meaningful components: trait, the characteristic of interest; context, 'factors related to place and circumstance that influence the subject's expression of [the trait]' (p. 1569); and perspective, which is characteristics of informants that impact their judgments. Similarly, other investigators have used factor-analytic strategies to derive latent variables capturing different aspects of informants' ratings. For example, Dumenci et al. (2011) created factors reflecting a higher-order externalizing trait generalized across raters, and lower-order traits reflecting source-specific variability. Finally, a number of research teams have used latent class analysis to identify groups of children differentiated by their behavior in specific contexts (e.g., behavior is displayed when interacting with a parent, with a stranger, or both; De Los Reyes et al., 2009) or as perceived by different informants (e.g., high ratings of problematic behavior given by mother only, teacher only, or both; Fergusson et al., 2009).

As a beginning step, the last approach may hold the most promise for case conceptualization. This strategy could be adapted for clinical use by identifying meaningful cut points on dimensions of interest as rated by a particular informant and using those to classify children. For example, children would be grouped as manifesting clinically significant ODD as identified by parent only, teacher only, or both (e.g., Drabick et al., 2007). Clinicians, fundamentally, have to make a dichotomous decision – treat or not – and given the evidence reviewed previously, children in all three groups would warrant intervention. However, what type of intervention, and how children could respond, might vary meaningfully and in unexpected ways if the differences between these groups are not limited to phenomenology. There is some evidence, for instance, that children with pervasive hyperactivity benefit more from treatment with stimulants than children with situational hyperactivity (Schachar & Tannock, 1993).

Clinicians are sensitive to the context in which symptoms occur (Pottick et al., 2007), and many will be incorporating this type of information into their conceptualizations already. Systematizing this process provides an opportunity to examine critically the potential clinical utility of such distinctions, allowing for further refinements. For example, it would be important to establish that there is predictive power with regard to treatment outcome associated with establishing categories based on patterning of ratings across informants. Preliminary evidence could be obtained by reanalyzing existing data to ascertain (a) whether it is possible to obtain consistently meaningful classifications of children into these groupings, and (b) their associations with correlates and outcomes, both normatively and in response to intervention. The existing diagnostic categories likely provide a useful starting point. Given the movement within the field to identifying core, underlying mechanisms of psychopathology (Insel et al., 2010), it may eventually be fruitful to examine inter-informant variability in more specific behavioral and emotional processes.

Although there is anticipated benefit to maintaining cross-informant patterning at the onset of treatment, considering each rater separately may be the most useful strategy for monitoring progress. Research suggests that raters' ability to report on behavior outside of their own setting is limited. For example, parental report on behavior at school shows markedly higher correlations with their ratings of behavior at home than with teacher report of behavior at school, and the converse is also true (De Nijs et al., 2004; Mitsis, McKay, Schulz, Newcorn, & Halperin, 2000). As such, report from an informant in one setting may not capture adequately functioning in a different context. Given the situation specificity of children's behavior, more generalized response to intervention may not always occur, making it important to collect data from an informant with first-hand knowledge of the setting of interest (De Los Reyes & Kazdin, 2009). One concern about using a source-specific approach is that it may yield lower quality measurement than strategies that combine information. There is some evidence to suggest, however, that reliability of source-specific ratings is comparable to a number of other data aggregation approaches, including both the 'and' and the 'or' rule for symptoms (Drabick et al., 2007; Jensen et al.,1995; Kraemer et al., 2003; Munkvold et al., 2009; Offord et al.,1996).

*Informants.* As the preceding review has made clear, there is substantial clinical utility associated with collecting information from both parents and teachers when making diagnostic decisions concerning ODD, CD, and ADHD (for additional evidence, see Owens & Hoza, 2003; Pelham et al., 2005; Loeber, Green, Lahey, & Stouthamer-Loeber, 1989). Most of this work has been conducted with school-aged children but there is evidence that teacher reports will also be useful for those who attend preschool (e.g., Murray et al., 2007). Obtaining self-report from children is also informative under some circumstances. Depending on the instrument used, young children may not be able to provide a reliable report (Frick & McMahon, 2008). For older children, however, self-report is a critical piece of the puzzle in the assessment of CD, likely due to the fact that

many of the behaviors occur in settings to which adults are not privy (Cantwell, Lewinsohn, Rohde, & Seeley, 1997; Jensen et al., 1999; Loeber et al., 1989), In contrast, children's self-report of ADHD symptoms is of limited value (Pelham et al., 2005), and there is debate concerning how much children's self-report of ODD symptoms contributes beyond parental report (Angold & Costello, 1996; Jensen et al., 1999; Loeber et al., 1989).

Thus far, the research reviewed provides information about a given class of informants, on average. One question with which clinicians must wrestle is whether there are conditions under which reports provided by a particular rater may not be credible (De Los Reyes et al., 2011b; Youngstrom et al., 2011). Given the reliance of clinicians and researchers on maternal report, there has been substantial interest in factors that may impact mothers' judgments, with much work focusing on whether maternal depression is associated with a tendency to over endorse disruptive behavior problems. This phenomenon has been demonstrated (e.g., Boyle & Pickles, 1997; Briggs-Gowan, Carter, & Schwab-Stone, 1996; but see Conrad & Hammen, 1989), but the magnitude of the bias may actually be quite small, indicating that there is still value in these reports (Youngstrom, Izard, & Ackerman, 1999). When considering teacher ratings, concern has been raised that there may be a systematic over-reporting of externalizing problems for minority children (e.g., Epstein et al., 2005); however, empirical support for this position is equivocal. Some studies are consistent with this hypothesis, (e.g., Sonuga-Barke, Minocha, Taylor, & Sandberg, 1993; see Lau et al., 2004; for review), but others are not (e.g., Chang & Sue, 2003; Epstein et al., 2005; Hosterman, DuPaul, & Jitendra, 2008), with evidence appearing stronger for disruptive behaviors than for ADHD. Some researchers have suggested that such biases may be due to a cultural mismatch between teachers, who, at least in the United States, are predominantly non-Hispanic white (Hosterman et al., 2008), and their students (Puig et al., 1999); however, data addressing this issue appear sparse and do not clearly indicate that congruence between teacher and student ethnicity will yield a more accurate accounting (see De Ramirez & Shapiro, 2005; Pigott & Cowen, 2000).

Even if reports by parents and teachers, on average, do not show evidence of substantial bias, clinicians will always confront individual cases in which they are concerned about the veracity of a report (e.g., the informant uses substances; Youngstrom et al., 2011). In recent research, Youngstrom et al. (2011) examined clinicians' perceptions of informants' credibility. Results indicated that informants seen as less credible did provide less valid information, but the authors concluded that these differences were not great enough to justify discarding the data. Although further work on this issue is needed, these findings suggest that it is rare that an informant's report is of no value, and that one fruitful direction for research would be the development of techniques to correct for systematic error in informants' reports, both at the individual and aggregate levels.

## 'Coming around again': application of advances in assessment to the refinement of conceptualizations of child psychopathology for DSM-5

Advancing understanding of how to obtain maximum benefit from informants' reports will increase the clinical utility of these instruments, but ultimately what is needed is greater understanding of the meaning of informant disagreement for conceptualizations of clinical phenomenology. Although inter-rater discrepancies in judgments of children's psychopathology have been viewed as a problem, these differences reflect meaningful variability in children's behavior and informants' perspectives across contexts. As such, the presence of informant disagreement provides an opportunity to advance theory and nosology in childhood psychopathology, which, in turn, should contribute to an increased understanding of developmental mechanisms. For example, there is growing evidence that genetic contributions to childhood psychopathology vary as a function of informant (e.g., Burt, 2009; Gizer et al., 2008). Such work suggests further unpacking informant discrepancies will advance clinical practice not only by enabling the development of more valid and efficient assessment techniques, but also by contributing to fundamental understanding of the etiology and maintenance of psychiatric disorders in childhood.

The critical next step for this line of research is to disentangle the relative contribution of (a) situational variability in behavior, and (b) rater-specific variables. To date, work on the situation specificity of children's behavior has been conducted along disparate lines from investigation of inter-rater discrepancies in evaluations of children's psychological symptomatology. The merging of these two traditions (e.g., De Los Reyes et al., 2009; Hartley et al., 2011) will be essential as researchers work to delineate the extent to which variability in informants' evaluations are driven by differences in the behavior of the child across contexts versus factors related to the informant, including both bias and varying perspectives resulting from the demands of a particular setting. Such work is already underway. For example, a recent investigation by Gomez (2007) used IRT to demonstrate that ADHD symptoms were perceived in a similar way by parents and teachers, suggesting that the low agreement between these informants was resulting from cross-setting differences in children's behavior. More research of this

type is needed to examine the generalizability of these findings to other symptom types, as well as to clinical samples.

It is also imperative that research move beyond the common practice of confounding informant and setting (i.e., using parent report to assess behavior at home and teacher report to assess behavior at school; Drabick, 2011), which often complicates interpretation of data due to the issue of shared method variance (see Costello et al., 1991) and provides only a crude measure of children's behavior across settings and situations. This decoupling can be achieved by systematically assessing differences in specific behaviors across situations directly, resulting in a clearer mapping of the roles of situational and informant factors in inter-rater discrepancies (De Los Reyes et al., 2009). This work will contribute to the continued development of theoretical models of developmental psychopathology. In this paper, we have suggested two possible ways to parse children's symptomatology: functional groupings, based on the situations in which symptoms occur, and source-specific categories, defined by the combinations of informants who have identified clinically significant syndromes or behaviors. Although related, these conceptualizations are not the same, and in order to determine which approach is more valid, it is necessary to separate children's behavior from the informant, so that the contributions of each may be analyzed.

Advances in contextualized measurement make it possible to answer these questions. Much work in this area has relied on intensive, naturalistic observations (e.g., Wright et al., 2011), which provide a rich behavioral sample, but are impractical, particularly in clinical settings. It is now clear that it is possible to capture reliable, clinically meaningful, contextual variability in behavior using interviews and rating scales (e.g., Dirks et al., 2007b; Wright & Zakriski, 2001), as well as brief, structured observational tasks (Wakschlag et al., 2008a). The increased feasibility of these approaches will allow researchers to conduct studies explicitly examining the associations between situational factors and symptomatology with a variety of samples and in an increased range of settings, providing significant opportunity to advance understanding of the role of situation- and setting-level factors in externalizing behavior problems. For example, Gray et al. (2011) utilized the contextualized measurement afforded by the DB-DOS to demonstrate that the pervasiveness of disruptive behavior may be less clinically informative for girls. Specifically, they showed that disruptive boys were disruptive during interactions with both parent and examiner, whereas disruptive girls showed high levels of disruptive behavior only when interacting with their parents. These findings suggest that a cross-contextual pervasiveness requirement may under identify clinically significant disruptive behavior in girls, information that could only be obtained through the use of standardized, contextually sensitive measures.

## Conclusions

The role of context in the development and maintenance of children's behavior problems has long been recognized by clinicians in their day-to-day work with individual children and their families. This knowledge, however, has not been widely integrated into measurement tools, nor into conceptualizations of psychopathology. Yet, there is increasing evidence that behavioral differences across settings and situations are reliable and meaningful, data that suggest that developing a more fine-grained understanding of the contextualized patterns of children's symptomatology will advance our knowledge of developmental psychopathology. As the field pushes toward DSM-5 there is an opportunity to consider how to strengthen the existing nosological framework. Considering the specific conditions under which symptomatology manifests, and measuring these contingencies systematically, may aid in the refinement of psychiatric phenotypes, work that may be necessary to push the boundaries of our knowledge of the etiology and maintenance of childhood psychiatric disorder.

Increased attention to the role of context in the expression of psychological symptoms should also translate into more precise assessment of clinical phenomena, ultimately bolstering the utility of our assessment approaches. For example, advances in contextualized measurement have helped, in part, to address the absence of developmental considerations that has characterized the disruptive syndromes (Wakschlag et al., 2010) by providing a more detailed framework by which to evaluate whether behaviors are clinically concerning or within normative bounds. Incorporation of contextual features could pay dividends for the creation of developmentally sensitive measures at other stages of childhood and adolescence, an issue that has received little attention (Carter, Gray, Baillargeon, & Wakschlag, in press).

Such focus on the utility of measurement approaches remains critically important. Given the enormous and growing strain on the mental health system, it is essential that assessment procedures be as streamlined as possible, with each approach and informant contributing substantially to diagnosis and treatment. The incremental validity of different techniques has received insufficient attention from researchers and we recommend that the bar be raised in regard to standards of evidence for inclusion of multiple methods and informants for treatment and prediction (see Hunsley & Meyer, 2003). There are data suggesting that briefer rating scales perform as well as lengthier interviews for some purposes, as well as substantial evidence indicating

that acquiring information from children's teachers about disruptive behavior syndromes and ADHD is worth the extra effort. Much work remains, however. Utility will be heavily influenced by developmental concerns, but little work has evaluated whether different methods are more informative during particular periods of childhood. Establishing treatment utility by determining the extent to which assessments contribute to outcomes in intervention will provide a strong case for their inclusion and will help to trim unnecessary procedures from assessment batteries. As the field advances and we continue to deepen our understanding of which assessment practices are most efficient, for whom, and when, the goal should not be the eradication of differences across informants and methods. Rather, these differences should be embraced, as they reflect meaningful information that could play an important role in clinical decision-making. Ultimately, further elucidation of their causes will yield significant theoretical dividends, enhancing both our measurement, and eventually, our intervention practices.

## Acknowledgements

The writing of this paper has been supported by NIMH grant R01MH082830 to Drs. Wakschlag and Briggs-Gowan and support to Dr. Wakschlag by the Walden & Jean Young Shaw Foundation. Dr. Dirks is grateful to Dr. Timothy Strauman for his comments on earlier versions of this manuscript, as well as to her student, Laura Bellhouse.

This review article was invited by the journal, for which the principal author has been offered a small honorarium payment towards personal expenses. The authors have declared that they have no competing or potential conflicts of interest (see footnote on p. 558).

## Correspondence to

Melanie A. Dirks, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, Canada, H3A1B1; Tel: 514 398 3856; Email: melanie.dirks@mcgill.ca

## Key points

- The tools and techniques used to assess developmental psychopathology must be consistent with theoretical models of the phenomena, and data yielded by advances in measurement should contribute to refinement of these conceptualizations.
- Children's behavior varies meaningfully across contexts, differences that, in combination with informants' perspectives, contribute to inter-rater discrepancies in symptom reports.
- Incorporating contextual features into measurement approaches (e.g., maintaining patterns of ratings across informants rather than collapsing them together) will contribute to conceptual understanding of psychopathology and enhance the clinical utility of assessment instruments.
- Clinical utility of methods and informants must be considered carefully, relative to the goal of the assessment, and the 'value added' of more intensive methods and additional informants must be demonstrated.

## References

Abikoff, H., Courtney, M., Pelham, W.E., & Koplewicz, H.S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*, 519–533.

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

Achenbach, T.M., & Rescorla, L.A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth and Families.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th edn, text rev.). Washington, DC: Author.

Angold, A., & Costello, E.J. (1996). The relative diagnostic utility of child and parent reports of oppositional defiant behaviors. *International Journal of Methods in Psychiatric Research*, *6*, 253–259.

Angold, A., & Costello, E.J. (2000). The Child and Adolescent Psychiatric Assessment (CAPA). *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*, 39–48.

Angold, A., & Costello, E.J. (2009). Nosology and measurement in child and adolescent psychiatry. *Journal of Child Psychology and Psychiatry*, *50*, 9–15.

Baillargeon, R.H., Boulerice, B., & Tremblay, R.E. (2001). Modeling interinformant agreement in the absence of a "gold standard". *Journal of Child Psychology and Psychiatry*, *42*, 463–473.

Boyle, M.H., Offord, D.R., Racine, Y.A., Szatmari, P., Sanford, M., & Fleming, J.E. (1997). Adequacy of interviews vs checklists for classifying childhood psychiatric disorder based on parent reports. *Archives of General Psychiatry*, *54*, 793–799.

Boyle, M.H., & Pickles, A. (1997). Influence of maternal depressive symptoms on ratings of childhood behavior. *Journal of Abnormal Child Psychology*, *25*, 399–412.

Brendgen, M., Vitaro, F., Boivin, M., Dionne, G., & Perusse, D. (2006). Examining genetic and environmental effects on reactive versus proactive aggression. *Developmental Psychology*, *42*, 1299–1312.

Briggs-Gowan, M.J., Carter, A.S., & Schwab-Stone, M. (1996). Discrepancies among mother, child, and teacher reports: Examining the contributions of maternal depression and anxiety. *Journal of Abnormal Child Psychology*, *24*, 749–765.

Burt, S.A. (2009). Rethinking environmental contributions to child and adolescent psychopathology: A meta-analysis of shared environmental influences. *Psychological Bulletin*, *135*, 608–637.

Burt, S.A., McGue, M., Krueger, R.F., & Iacono, W.G. (2005). Sources of covariation among the child-externalizing disor-

ders: Informant effects and the shared environment. *Psychological Medicine*, *35*, 1133–1144.

Cantwell, D.P., Lewinsohn, P.M., Rohde, P., & Seeley, J.R. (1997). Correspondence between adolescent report and parent report of psychiatric diagnostic data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*, 610–619.

Carter, A.S., Gray, S.A.O., Baillargeon, R.H., & Wakschlag, L.S. (in press). A multidimensional approach to disruptive behaviors: Informing lifespan research from an early childhood perspective. In P.H. Tolan, & B.L. Leventhal (Eds.), *Brain Research Foundation Symposium Series Advances in development and psychopathology: Volume 1, Disruptive behavior disorders*. New York: Springer.

Cella, D., Gershon, R., Lai, J.S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(Supplement 1), 133–141.

Chang, D.F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology*, *71*, 235–242.

Cole, P.M., Martin, S.E., & Dennis, T.A. (2004). Emotion regulation as a scientific construct: Methodological challenges and directions for child development research. *Child Development*, *75*, 317–333.

Cole, D.A., Martin, J.M., Powers, B., & Truglio, R. (1996). Modeling causal relations between academic and social competence and depression: A multi-trait, multi-method longitudinal study of children. *Journal of Abnormal Psychology*, *105*, 258–270.

Collishaw, S., Goodman, R., Ford, T., Rabe-Hesketh, S., & Pickles, A. (2009). How far are associations between child, family and community factors and child psychopathology informant-specific and informant-general? *Journal of Child Psychology and Psychiatry*, *50*, 571–580.

Conrad, M., & Hammen, C. (1989). Role of maternal depression in perceptions of child maladjustment. *Journal of Consulting and Clinical Psychology*, *57*, 663–667.

Costello, E.J., Egger, H., & Angold, A. (2005). 10-year research update review: The epidemiology of child and adolescent psychiatric disorders: I. Methods and public health burden. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 972–986.

Costello, E.J., Loeber, R., & Stouthamer-Loeber, M. (1991). Pervasive and situational hyperactivity – Confounding effect of informant: A research note. *Journal of Child Psychology and Psychiatry*, *32*, 367–376.

Crick, N.R., & Dodge, K.A. (1996). Social information-processing mechanisms in reactive and proactive aggression. *Child Development*, *67*, 993–1002.

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *40*, 1–9.

De Los Reyes, A., Alfano, C.A., & Beidel, D.C. (2010). The relations among measurements of informant discrepancies within a multisite trial of treatments for childhood social phobia. *Journal of Abnormal Child Psychology*, *38*, 395–404.

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

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

De Los Reyes, A., & Kazdin, A.E. (2009). Identifying evidence-based interventions for children and adolescents using the range of possible changes model: A meta-analytic illustration. *Behavior Modification*, *33*, 583–617.

De Los Reyes, A., Youngstrom, E.A., Pabón, S.C., Youngstrom, J.K., Feeny, N.C., & Findling, R.L. (2011a). Internal consistency and associated characteristics of informant discrepancies in clinic referred youths age 11 to 17 years. *Journal of Clinical Child and Adolescent Psychology*, *40*, 36–53.

De Los Reyes, A., Youngstrom, E.A., Swan, A.J., Youngstrom, J.K., Feeny, N.C., & Findling, R.L. (2011b). Informant discrepancies in clinical reports of youths' and interviewers' impressions of the reliability of informants. *Journal of Child and Adolescent Psychopharmacology*, *21*, 417–424.

De Nijs, P.F.A., Ferdinand, R.F., de Bruin, E.I., Dekker, M.C.J., van Duijn, C.M., & Verhulst, D.C. (2004). Attention-deficit/hyperactivity disorder (ADHD): Parents' judgment about school, teachers' judgment about home. *European Child and Adolescent Psychiatry*, *13*, 315–320.

De Ramirez, R.D., & Shapiro, E.S. (2005). Effects of student ethnicity on judgments of ADHD symptoms among Hispanic and White teachers. *School Psychology Quarterly*, *20*, 268–287.

Dirks, M.A., & Boyle, M.H. (2010). The comparability of mother-report structured interviews and checklists for the quantification of youth externalizing symptoms. *Journal of Child Psychology and Psychiatry*, *51*, 1040–1049.

Dirks, M.A., Boyle, M.H., & Georgiades, K. (2011a). Psychological symptoms in youth and later socioeconomic functioning: Do associations vary by informant? *Journal of Clinical Child and Adolescent Psychology*, *40*, 10–22.

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.

Dirks, M.A., Treat, T.A., & Weersing, V.R. (2010). The judge specificity of evaluations of youth social behaviour: The case of peer provocation. *Social Development*, *19*, 736–757.

Dirks, M.A., Treat, T.A., & Weersing, V.R. (2011b). The latent structure of youth responses to peer provocation. *Journal of Psychopathology and Behavioral Assessment*, *33*, 58–68.

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

Drabick, D.A.G. (2009). Can a developmental psychopathology perspective facilitate a paradigm shift toward a mixed categorical-dimensional classification system? *Clinical Psychology: Science and Practice*, *16*, 41–49.

Drabick, D.A.G. (2011). Report to the DSM-5 ADHD and Disruptive Behaviors Working Group on Oppositional Defiant Disorder. Unpublished manuscript. Available from: http://www.dsm5.org/ProposedRevision/Pages/proposedrevision.aspx?rid=106# [last accessed 5 November 2011].

Drabick, D.A.G., Bubier, J., Chen, D., Price, J., & Lanza, I.H. (2011). Source-specific oppositional defiant disorder among inner-city children: Prospective prediction and moderation. *Journal of Clinical Child and Adolescent Psychology*, *40*, 23–35.

Drabick, D.A.G., Gadow, K.D., & Loney, J. (2007). Source-specific oppositional defiant disorder: Comorbidity and risk factors in referred elementary schoolboys. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*, 92–101.

Dumenci, L., Achenbach, T.M., & Windle, M. (2011). Measuring context-specific and cross-contextual components of hierarchical constructs. *Journal of Psychopathology and Behavioral Assessment*, *33*, 3–10.

DuPaul, G.J., & Barkley, R.A. (1992). Situational variability of attention problems: Psychometric properties of the revised home and school situations questionnaires. *Journal of Clinical Child Psychology*, *21*, 178–188.

Edelbrock, C., & Costello, A.J. (1988). Convergence between statistically derived behavior problem syndromes and child

psychiatric diagnoses. *Journal of Abnormal Child Psychology*, *16*, 219–231.

Epstein, J.N., Willoughby, M., Valencia, E.Y., Toney, S.T., Abikoff, H.B., Arnold, L.E., & Hinshaw, S.P. (2005). The role of children's ethnicity in the relationship between teacher ratings of attention-deficit/hyperactivity disorder and observed classroom behavior. *Journal of Consulting and Clinical Psychology*, *73*, 424–434.

Fergusson, D.M., Boden, J.M., & Horwood, L.J. (2009). Situational and generalised conduct problems and later life outcomes: Evidence from a New Zealand birth cohort. *Journal of Child Psychology and Psychiatry*, *50*, 1084–1092.

Frick, P.J., & McMahon, R.J. (2008). Child and adolescent conduct problems. In J. Hunsley, & E.J. Mash, (Eds.), *A guide to assessments that work* (pp. 41–68). New York: Oxford.

Frick, P.J., Lahey, B.B., Applegate, B., Kerdyck, L., Ollendick, T., Hynd, G.W, … & Waldman, I. (1994). DSM-IV field trials for the disruptive behavior disorders: Symptom utility estimates. *Journal of the American Academy of Child and Adolescent Psychiatry*, *33*, 529–539.

Garb, H. (2007). Computer-administered interviews and rating scales. *Psychological Assessment*, *19*, 4–13.

Gardner, F. (2000). Methodological issues in the direct observation of parent-child interaction: Do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review*, *3*, 185–198.

Gelhorn, H., Hartman, C., Sakai, J., Mikulich-Gilbertson, S., Stallings, M., Young, S., …. Crowley, T. (2009). An item-response theory analysis of DSM-IV conduct disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *48*, 42–50.

Ginsburg, B.E., Werick, T.M., Escobar, J.I., Kugelmass, S., Treanor, J.J., & Wendtland, L. (1996). Molecular genetics of psychopathologies: A search for simple answers to complex problems. *Behavior Genetics*, *26*, 325–333.

Gizer, I.R., Waldman, I.D., Abramowitz, A., Barr, C., Feng, Y., Wigg, K.G., …. Rowe, D.C. (2008). Relations between multi-informant assessments of ADHD symptoms, DAT1, and DRD4. *Journal of Abnormal Psychology*, *117*, 869–880.

Gomez, R. (2007). Australian parent and teacher ratings of the DSM-IV ADHD symptoms: Differential symptom functioning, and parent-teacher agreement and differences. *Journal of Attention Disorders*, *11*, 17–27.

Gomez, R., Burns, G.L., Walsh, J.A., & de Moura, M.A. (2003). A multitrait-multisource confirmatory factor analytic approach to the construct validity of ADHD rating scales. *Psychological Assessment*, *15*, 3–16.

Gray, S., Carter, A., Briggs-Gowan, M.J., Hill, C., Danis, B., & Wakschlag, L.S. (2011). Preschool children's observed disruptive behavior: Variations across sex, interactional context, and severity of disruptive behavior. Manuscript submitted for publication.

Grayson, P., & Carlson, G.A. (1991). The utility of a DSM-III-R-based checklist in screening child psychiatric patients. *Journal of the American Academy of Child & Adolescent Psychiatry*, *30*, 669–673.

Gresham, F.M., & Elliott, S.N. (2008). *Social skills improvement system: Rating scales*. Bloomington, MN: Pearson Assessments.

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., Leisen, M., & Blaine, D.D. (1997). Design of individualized behavioral treatment programs using functional analytic clinical case models. *Psychological Assessment*, *9*, 334–348.

Ho, T.P., Luk, E.S.L., Leung, P.W.L., Taylor, E., Lieh-Mak, F., & Bacon-Shone, J. (1996). Situational versus pervasive hyperactivity in a community sample. *Psychological Medicine*, *26*, 308–321.

Holmbeck, G.N., Li, S.T., Schurman, J.V., Friedman, D., & Coakley, R.M. (2002). Collecting and managing multisource and multimethod data in studies of pediatric populations. *Journal of Pediatric Psychology*, *27*, 5–18.

Hosterman, S.J., DuPaul, G.J., & Jitendra, A.K. (2008). Teacher ratings of ADHD symptoms in ethnic minority students: Bias or behavioral difference? *School Psychology Quarterly*, *23*, 418–435.

Hudziak, J.J., Achenbach, T.M., Althoff, R.R., & Pine, D.S. (2007). A dimensional approach to developmental psychopathology. *International Journal of Methods in Psychiatric Research*, *16*(S1), S16–S23.

Hunsley, J., & Mash, E.J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29–51.

Hunsley, J., & Mash, E.J. (Eds.), (2008). *A guide to assessments that work*. New York: Oxford.

Hunsley, J., & Meyer, G.J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446–455.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., … Wang, P. (2010). Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*, 748–750.

Jensen, P., Roper, M., Fisher, P., Piacentini, J., Canino, G., Richters, J., … Schwab-Stone, M. (1995). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). *Archives of General Psychiatry*, *52*, 61–71.

Jensen, P.S., Rubio-Stipec, M., Canino, G., Bird, H.R., Dulcan, M.K., Schwab-Stone, M.E., & Lahey, B.B. (1999). Parent and child contributions to diagnosis of mental disorder: Are both informants always necessary? *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*, 1569–1579.

Jensen, A.L., & Weisz, J.R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*, *70*, 158–168.

Jensen-Doss, A., & Hawley, K. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child and Adolescent Psychology*, *39*, 885–896.

Johnston, C., & Mah, J.W.T. (2008). Child attention-deficit/hyperactivity disorder. In J. Hunsley, & E.J. Mash (Eds.), *A guide to assessments that work* (pp. 17–40). New York: Oxford.

Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment*, *15*, 496–507.

Kendell, R., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *The American Journal of Psychiatry*, *160*, 4–12.

Kendler, K.S. (2005). Toward a philosophical structure for psychiatry. *American Journal of Psychiatry*, *162*, 433–440.

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

Laird, R.D., & Weems, C.F. (2011). The equivalence of regression models using difference scores and models using separate scores for each informant: Implications for the study of informant discrepancies. *Psychological Assessment*, *23*, 388–397.

Lau, A.S., Garland, A.F., Yeh, M., McCabe, K.M., Wood, P.A., & Hough, R.L. (2004). Race/ethnicity and inter-informant agreement in assessing adolescent psychopathology. *Journal of Emotional and Behavioral Disorders*, *12*, 145–156.

Loeber, R., Green, S.M., Lahey, B.B., & Stouthamer-Loeber, M. (1989). Optimal informants on childhood disruptive behaviors. *Development and Psychopathology*, *1*, 317–337.

Lord, C., Risi, S., Lambrecht, L., Cook, E., Leventhal, B., DiLavore, P., Pickles, A., & Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*, 205–223.

Lothmann, C., Holmes, E.A., Chan, S.W., & Lau, J.Y. (2011). Cognitive bias modification training in adolescents: Effects on interpretation biases and mood. *Journal of Child Psychology and Psychiatry*, *52*, 24–32.

Mannuzza, S., Klein, R.G., & Moulton, J.L. (2002). Young adult outcome of children with "situational" hyperactivity: A prospective, controlled follow-up study. *Journal of Abnormal Child Psychology*, *30*, 191–198.

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 behavior in peer-peer context. *Journal of Child Psychology and Psychiatry*, *42*, 413–420.

McConaughy, S.H., Harder, V.S., Antshel, K.M., Gordon, M., Eiraldi, R., & Dumenci, L. (2010). Incremental validity of test session and classroom observations in a multimethod assessment of attention deficit/hyperactivity disorder. *Journal of Clinical Child and Adolescent Psychology*, *39*, 650–666.

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

McFall, R.M. (2005). Theory and utility-key themes in evidence-based assessment: Comment on the special section. *Psychological Assessment*, *17*, 312–323.

McFall, R.M., & Townsend, J.T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment*, *10*, 316–330.

Mitsis, E.M., McKay, K.E., Schulz, K.P., Newcorn, J.H., & Halperin, J.M. (2000). Parent-teacher concordance for DSM-IV attention deficit/hyperactivity disorder in a clinic-referred sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*, 308–313.

Munkvold, L., Lundervold, A., Lie, S.A., & Manger, T. (2009). Should there be separate parent and teacher-based categories of ODD? Evidence from a general population. *Journal of Child Psychology and Psychiatry*, *50*, 1264–1272.

Murray, D.W., Kollins, S.H., Hardy, K.K., Abikoff, H.B., Swanson, J.M., Cunningham, C.,.... Chuang, S.Z. (2007). Parent versus teacher ratings of attention-deficit/hyperactivity disorder symptoms in the Preschoolers with Attention-Deficit/Hyperactivity Disorder Treatment Study (PATS). *Journal of Child and Adolescent Psychopharmacology*, *17*, 605–619.

Offord, D.R., Boyle, M.H., Racine, Y., Szatmari, P., Fleming, J.E., Sanford, M., & Lipman, E.L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 1078–1085.

Owens, J., & Hoza, B. (2003). Diagnostic utility of DSM-IV-TR symptoms in the prediction of DSM-IV-TR ADHD subtypes and ODD. *Journal of Attention Disorders*, *7*, 11–27.

Pelham, W.R., Fabiano, G.A., & Massetti, G.M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 449–476.

Perren, S., Von Wyl, A., Stadelmann, S., Burgin, D., & Von Klitzing, K. (2006). Associations between behavioral/emotional difficulties in kindergarten children and the quality of their peer relationships. *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 867–876.

Pigott, R.L., & Cowen, E.L. (2000). Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. *Journal of School Psychology*, *38*, 177–196.

Pottick, K.J., Kirk, S.A., Hsieh, D.K., & Tian, X. (2007). Judging mental disorder: Effects of client, clinician, and contextual differences. *Journal of Consulting and Clinical Psychology*, *75*, 1–8.

Power, T.J., Costigan, T.E., Leff, S.S., Eiraldi, R.B., & Landau, S. (2001). Assessing ADHD across settings: Contributions of behavioral assessment to categorical decision making. *Journal of Clinical Child Psychology*, *30*, 399–412.

Puig, M., Lambert, M.C., Rowan, G.T., Winfrey, T., Lyubansky, M., Hannah, S.D., & Hill, M.F. (1999). Behavioral and emotional problems among Jamaican and African-American children, ages 6 to 11: Teacher reports versus direct observations. *Journal of Emotional and Behavioral Disorders*, *7*, 240–250.

Rapoport, J.L., Donnelly, M., Zametkin, A., & Carrougher, J. (1986). 'Situational hyperactivity' in a U.S. clinical setting. *Journal of Child Psychology and Psychiatry*, *27*, 639–646.

Reise, S.P., & Waller, N.G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Rettew, D.C., Lynch, A.D., Achenbach, T.M., Dumenci, L., & Ivanova, M.Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research*, *18*, 169–184.

Rettew, D.C., van Oort, F.V.A., Verhulst, F.C., Buitelaar, J.K., Ormel, J., Hartman, C., .... Hudziak, J.J. (2011). When parent and teacher ratings don't agree: The Tracking Adolescents' Individual Lives Survey (TRAILS). *Journal of Child and Adolescent Psychopharmacology*, *21*, 389–397.

Rose, A.J., & Asher, S.R. (1999). Children's goals and strategies in response to conflicts within a friendship. *Developmental Psychology*, *35*, 69–79.

Rowe, D.C., & Kandel, D. (1997). In the eye of the beholder? Parental ratings of externalizing and internalizing symptoms *Journal of Abnormal Child Psychology*, *25*, 265–275.

Sanislow, C.A., Pine, D.S., Quinn, K.J., Kozak, M.J., Garvey, M.A., Heinssen, R.K., Wang, P.S., & Cuthbert, B.N. (2010). Developing constructs for psychopathology research: Research domain criteria. *Journal of Abnormal Psychology*, *119*, 631–639.

Schachar, R., Rutter, M., & Smith, A. (1981). The characteristics of situationally and pervasively hyperactive children: Implications for syndrome definition. *Journal of Child Psychology and Psychiatry*, *22*, 375–392.

Schachar, R., & Tannock, R. (1993). Childhood hyperactivity and psychostimulants: A review of extended treatment studies. *Journal of Child and Adolescent Psychopharmacology*, *3*, 81–97.

Schachar, R.J., Tannock, R., & Logan, G. (1993). Inhibitory control, impulsiveness, and attention deficit hyperactivity disorder. *Clinical Psychology Review*, *13*, 721–739.

Silverman, W.K., & Ollendick, T.H. (2005). Evidence-based assessment of anxiety and its disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 380–411.

Simonoff, E., Pickles, A., Hervas, A., Silberg, J.L., Rutter, M., & Eaves, L. (1998). Genetic influences on childhood hyperactivity: Contrast effects imply parental rating bias, not sibling interaction. *Psychological Medicine*, *28*, 825–837.

Sonuga-Barke, E.J.S., Minocha, K., Taylor, E.A., & Sandberg, S. (1993). Inter-ethnic bias in teachers' ratings of childhood hyperactivity. *British Journal of Developmental Psychology*, *11*, 187–200.

Tripp, G.G., & Luk, S.L. (1997). The identification of pervasive hyperactivity: Is clinic observation necessary? *Journal of Child Psychology and Psychiatry*, *38*, 219–234.

Van Dulmen, M.H.M., & Egeland, B. (2011). Analyzing multiple informant data on child and adolescent behavior problems: Predictive validity and comparison of aggregation procedure. *International Journal of Behavioral Development*, *35*, 84–92.

Wakschlag, L.S., Briggs-Gowan, M.J., Carter, A.S., Hill, C., Danis, B., Keenan, K., … Leventhal, B.L. (2007). A developmental framework for distinguishing disruptive behavior from normative misbehavior in preschool children. *Journal of Child Psychology and Psychiatry*, *48*, 976–987.

Wakschlag, L.S., Briggs-Gowan, M., Choi, S., Hullsiek, H., Burns, J., McCarthy, K., … & Carter, A.S. (2008). *Defining the developmental parameters of temper loss in early childhood: Implications for developmental psychopathology*, Manuscript submitted for publication.

Wakschlag, L.S., Briggs-Gowan, M.J., Hill, C., Danis, B., Leventhal, B.L., Keenan, K., … Carter, A.S. (2008a). Observational assessment of preschool disruptive behavior, Part II: Validity of the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS). *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*, 632–641.

Wakschlag, L.S., Hill, C., Carter, A.S., Danis, B., Egger, H.L., Keenan, K., … Briggs-Gowan, M.J. (2008b). Observational assessment of preschool disruptive behavior, Part I: Reliability of the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS). *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*, 622–631.

Wakschlag, L.S., Tolan, P.H., & Leventhal, B.L. (2010). Research review: 'Ain't misbehavin': Towards a developmentally-specified nosology for preschool disruptive behavior. *Journal of Child Psychology and Psychiatry*, *51*, 3–22.

Weisz, J.R., McCarty, C.A., Eastman, K.L., Chaiyasit, W., & Sunwanlert, S. (1997). Developmental psychopathology and culture: Ten lessons from Thailand. In S. Luthar, J. Burack, D. Cicchetti, & J. Weisz (Eds.), *Developmental psychopathology: Perspectives on adjustment, risk, and disorder* (pp. 568–592). New York: Cambridge University Press.

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., 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.

Youngstrom, E., Izard, C., & Ackerman, B. (1999). Dysphoria-related bias in maternal ratings of children. *Journal of Consulting and Clinical Psychology*, *67*, 905–916.

Youngstrom, E.A., Youngstrom, J.K., Freeman, A.J., De Los Reyes, A., Feeny, N.C., & Findling, R.L. (2011). Informants are not all equal: Predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology*, *21*, 407–415.

Accepted for publication: 18 January 2012
Published online: 24 February 2012