JOURNAL OF CHILD AND ADOLESCENT PSYCHOPHARMACOLOGY
Volume 21, Number 5, 2011

Pp. 407–415
DOI: 10.1089/cap.2011.0032

Informants Are Not All Equal: Predictors and Correlates of Clinician Judgments About Caregiver and Youth Credibility

Eric A. Youngstrom, Ph.D.,[1] Jennifer Kogos Youngstrom, Ph.D.,[1] Andrew J. Freeman, M.A.,[1] Andres De Los Reyes, Ph.D.,[2] Norah C. Feeny, Ph.D.,[3] and Robert L. Findling, M.D.[3,4]

Abstract

Objective: The objectives of this study were to examine how often clinicians judged youths or caregivers to not be credible informants, to identify the associated features of youth or caregiver credibility, and to examine credibility's impact on the validity of mood and behavior checklists.
Background: Clinicians often have the experience of talking to a parent or a youth and judging that the credibility of the information offered is unusually poor. Little is known about the correlates of poor credibility or about the extent to which credibility changes the validity of commonly used checklists.
Methods: Interviewers rated the credibility of 646 youths aged 5–18 and their primary caregivers after completing a Kiddie Schedule for Affective Disorders and Schizophrenia. Ratings and diagnoses were blind to the behavior checklists completed by caregivers, youths, and teachers. A subset of youths also had intelligent quotient tests and behavioral observations available.
Results: Caregivers were perceived as more credible on average than youths, though this dropped sharply with adolescents. Caregiver credibility was higher for better functioning families, more credible youths, younger youths, and more educated caregivers; it was unrelated to caregiver mood symptoms or being the mother. Youth credibility was strongly connected to age, cognitive ability, caregiver credibility, and independent observations of youth behavior. Credibility ratings markedly altered the validity of checklists compared with interview ratings, diagnoses, or cross-informant criteria.
Conclusion: Clinicians' judgments about informant credibility are associated with different characteristics for youths versus caregivers, though youth age is important to both. Credibility affects the validity of information from checklists measured against several different independent criteria.

Introduction

Clinicians often have the experience of talking to a parent or a youth and judging that the credibility of the information offered is unusually poor. This could be due to factors such as impaired cognitive functioning, extreme demoralization or distress (Sellbom and Ben-Porath 2005), substance use issues (cf. Youngstrom et al. 2000), cultural differences in the conceptualization of problems (Chavez et al. 2010; Gonzalez et al. 2011), malingering (Henry et al. 2009), or multiple other considerations (Garb 1997; Groth-Marnat 1999). Clinicians may gauge these various issues in deciding how much weight to give to the perspectives of the different informants. Much research has documented the challenges clinicians face when attempting to adjust their formulations based on moderating circumstances, with a consistent theme being that clinical judgments often introduce the potential for bias in decision making (Meehl 1954; Dawes et al. 1989; Arkes 1991; Garb 1998). Credibility is distinct from the well-established concepts of ''reliability'' and ''validity.'' In broad terms, *reliability* refers to the reproducibility of scores, most often across variations of item, time, or rater (Kraemer et al. 2005). *Validity* describes the extent to which the score reflects the intended construct (Kraemer et al. 2005). We define *credibility* as a clinical judgment about the probable accuracy or bias of information from a particular source. Credibility is definitely linked to reliability and validity, but can be viewed as distinct from the two. Specifically, inconsistent reporting would reduce reliability, validity, and credibility, but it is possible to have highly reliable reports that have low credibility (as would occur when an informant has a clear agenda and always reports in ways consistent with those ulterior motives). Alternatively, because one derives credibility ratings based on clinical judgments, a person can provide a valid report that has low credibility if that person provides accurate information that a clinician decides arises from improbable circumstances.

To the clinician, perceptions of credibility are not so much about the ''average'' parent or the ''typical'' youth; they are observations

[1]Department of Psychology, University of North Carolina, Chapel Hill, North Carolina.
[2]Department of Psychology, University of Maryland, College Park, Maryland.
[3]Department of Psychology, Case Western Reserve University, Cleveland, Ohio.
[4]Department of Psychiatry, Case Western Reserve University School of Medicine, Cleveland, Ohio.

about an individual person. These judgments sometimes are based on a single, decisive piece of information, such as when a person appears intoxicated during the interview, and other times, the perception about credibility is based on a constellation of factors. To date, little or no work on cross-informant agreement has examined the extent to which clinicians' judgments about informant credibility might be associated with the validity of information derived from mood and behavior checklists. Two dominant approaches have been used to conceptualize differences in reported problems across informants: situational specificity and picking the best average informant (De Los Reyes and Kazdin 2005; De Los Reyes 2011). Both are described briefly, along with reasons why neither is fully satisfactory in clinical practice.

The ''situational specificity'' hypothesis has argued that each person is providing accurate information about behaviors that are specific to different situations (Achenbach 1995; Kraemer et al. 2003; De Los Reyes and Kazdin 2005). For decades, it has been well established that caregivers, youths, and teachers agree with each other at only modest levels when describing youth mood and behavior (Achenbach et al. 1987; Achenbach and Rescorla 2001; Reynolds and Kamphaus 2004). Information provided by each person typically meets high standards for internal consistency reliability, retest stability, and various aspects of validity (Achenbach and Rescorla 2001)—yet agreement remains modest. Teachers may provide accurate descriptions of behavior in the classroom, which could be different from parents' accurate description of behaviors at home, for example (Kraemer et al. 2003; Hudziak et al. 2005). Situational specificity is a widely accepted model (Sherman et al. 2010). Recent observational work in both laboratory and clinic settings supports it (De Los Reyes et al. 2009, Hartley et al. 2011), and the situational specificity model supports the common recommendation for clinicians to gather information from multiple sources (Mash and Hunsley 2005). However, this model creates challenges for clinicians when the information from different informants appears to disagree—especially when the options being considered for intervention are more global and cannot be adjusted for particular situations. A clinician choosing whether or not to prescribe a medication cannot specify that the atypical antipsychotic or mood stabilizer only affects the person at school, for example. The decision whether or not to medicate or initiate other forms of treatment becomes complicated when one person appears to deny a problem that another person reports is present (see Carlson and Blader, 2011).

The second approach to resolving disagreement has relied on identifying which is the most valid source of information for a particular diagnosis or problem, on average. When one source of information is clearly more valid than another, then the clinician can choose to ignore the less valid piece of information or even streamline the assessment process by not gathering information with lower validity. In fact, it is possible to dilute the validity of the information gathered by adding new results or scores with lower validity, yielding more expensive yet less accurate clinical decisions (Kraemer 1992). The ''pick the best'' strategy has led to discounting self-report by youths as a source of information about attention problems (Jensen et al. 1999) or giving greater credence to self-report about internalizing problems (Loeber et al. 1989). Conversely, the weight of evidence suggests the opposite about teacher report: teacher checklist ratings tend to be valid about attention problems (Barkley 1998; DuPaul et al. 1998) and much less so about internalizing disorders (Epkins 1995; Hazell et al. 1999; Youngstrom et al. 2004b; Youngstrom et al. 2008).

Problems with the ''pick the best'' approach include the fact that problems only reported by either the parent or the youth are still often associated with considerable impairment (Bird et al. 1992; Jensen et al. 1999; Youngstrom et al. 2003), as well as the possibility that there might be incremental value in combining different perspectives on youth functioning (e.g., Carlson and Youngstrom 2003). Differences between informants may also be more nuanced, playing out at the level of differing interpretations of specific items or behaviors rather than entire scales or domains (e.g., Freeman et al., 2011). A pragmatic concern is that the ''best'' informant based on the literature may not be available in practice. Much more research has focused on the validity of mothers as informants about child behavior as opposed to fathers or other caregivers (cf. Phares 1992, 1997). How should a clinician proceed when the youth in question is young and there is no mother involved, or the parent's perspective is clearly compromised by psychological factors or extrinsic considerations such as legal or custody proceedings?

A fundamental limitation of both the ''situational specificity'' and the ''pick the best'' algorithms is that they focus on group data, evaluating typical or average validity. These summary statistics are important starting points, but there may be huge individual differences hidden within the general trends. For example, a particularly insightful youth might have a clear and informative perspective about their own attention problems, whereas a caregiver who is dealing with serious impairments, or a foster parent who has only known the youth for a few months, will have much less basis for their perceptions—and correspondingly lower validity than usual.

Is there a way to capitalize on clinical judgments about the credibility of individual informants that supports diagnostic formulation and treatment planning? Can we do better than current strategies of saying that disagreement is due to situationally specific changes in behavior or picking a single best informant for each target issue? How valid would global judgments about credibility be? Would accounting for credibility lead to appreciable changes in the reliability or validity of information collected from each person?

The goals of the present study include (a) examining the relative frequency with which clinicians judged the information they received from youths and caregivers during an interview to be credible, (b) identifying factors associated with credibility of youth or caregiver report, and (c) examining whether judgments of credibility were associated with significant changes in the diagnostic or cross-informant validity of information from mood and behavior checklists.

Methods

Procedure

All procedures were reviewed and approved by the institutional review boards of University Hospitals of Cleveland, Applewood Centers, and the University of North Carolina at Chapel Hill. Participants were recruited from a consecutive case series seeking outpatient evaluation at either the largest community mental health center providing services to children and families in the state of Ohio or a neighboring academic medical center. A total of 646 youths ranging from age 5 to 18 years presented for an outpatient evaluation with their primary caregiver and completed a semi-structured diagnostic interview with highly trained raters. The interviewer rated the credibility of information received from the youth and caregiver as ''poor,'' ''fair,'' or ''good'' after talking with each informant. Caregivers also completed rating scales about

their youth's mood symptoms. Youths aged 11 and older completed the same rating scales about themselves, and teachers completed checklists about a subset of participants. Caregivers also reported their own current mood symptoms on two questionnaires. Diagnoses and credibility ratings were made blind to the scores on any mood and behavior checklists, which were gathered by a second research assistant. The caregiver completed the interview first when youths were younger than age 11. When the youth was older, the family was given their preference about interview order; 90% of families elected to have the caregiver complete the interview first.

Measures

Youth diagnoses and mood severity. A consensus meeting assigned youth diagnoses by reviewing the results of a Kiddie Schedule for Affective Disorders and Schizophrenia (KSADS) interview conducted by a highly trained rater ($K > 0.85$ at the item level for each of five training and five certification interviews) using the KSADS-Present and Lifetime version (Kaufman et al. 1997) with the Washington University mood disorders modules (Geller et al. 2001). The same interview provided the basis for scoring the severity of the youth's manic and depressive symptoms (Axelson et al. 2003). The interviewer met with the caregiver and the youth sequentially and reinterviewed each as necessary to use clinical judgment to resolve reporting discrepancies. A licensed clinical psychologist reviewed the KSADS findings in person with the interviewer and synthesized them with additional information about developmental history, treatment history, and family psychiatric history (Spitzer 1983). Diagnoses followed strict *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition, Text Revision (DSM-IV-TR), criteria (American Psychiatric Association 2000).

Credibility ratings. At the conclusion of the KSADS interviews, the interviewer also rated the credibility of the caregiver and the youth at the end of the interview day. The instructions were ''Reliability of information: 2 = good, 1 = fair, 0 = poor.'' Credibility scores were a global, subjective rating based on the clinical judgment of the interviewer.

Family characteristics. Age, race, and gender were primary demographic characteristics. The caregiver's relationship to the youth, their education level, their self-reported income, and their occupational status, along with the number of children in the home comprised the family-level demographic variables of interest for this study. Two additional sources of information rated the overall functioning of the family: the KSADS interviewer completed the Global Family Environment Scale (GFES; Rey et al. 1997), scoring each family on a scale from 1 to 90, with higher scores indicating better functioning. The primary caregiver completed the Family Assessment Device (FAD) (Byles et al. 1988) as part of the questionnaire packet. FAD total score provided a measure of overall discord in the family, combining information about poor communication, problem solving, and general functioning, with alpha = 0.91 in this sample.

Youth behavior problems. Caregivers completed the 2001 version of the Child Behavior Checklist (CBCL; Achenbach and Rescorla 2001). Youths aged 11 years and older ($n = 349$) also completed the self-report version of the same scale, and a subset of teachers ($n = 249$) also completed the Teacher Report Form. Teacher data were not gathered during the summer or first 6 weeks of the new school year, even though families continued to enroll in the project, because teachers were unavailable when schools were closed and less familiar with the youth's behavior at the beginning of the academic year. The present analyses focused on the Externalizing and Internalizing Problems Broad-Band *T*-scores and the Attention Problems Clinical Scale *T*-score, as these are the three domains that have received the most study in the cross-informant research literature.

Youth mood ratings. Primary caregivers also completed the Parent General Behavior Inventory (PGBI; Youngstrom et al. 2001) as a rating of the youth's depressed, hypomanic, and mixed mood symptoms. Youths 11 years and older also completed the self-report version (adolescent GBI), which has shown good psychometric properties within this age group (Danielson et al. 2003). Present analyses concentrated on the depression and hypomanic/biphasic scores, each with alpha > 0.94 in this sample.

Caregiver mood ratings. Primary caregivers also completed a Beck Depression Inventory (Beck and Steer 1987) and a Mood Disorder Questionnaire (Hirschfeld et al. 2000) as measures of their own depressive and hypomanic or manic symptoms.

Youth cognitive ability. A subset of youths ($n = 127$) completed cognitive ability testing as part of their clinical evaluation and had a global score available as part of the review of the treatment history. Present analyses used the global score, scaled with $M = 100$ and SD = 15. For 79% of cases, scores were based on the Wechsler Abbreviated Scales of Intelligence (The Psychological Corporation 1999), 12% based on the Peabody Picture Vocabulary Test–Third Edition (Dunn and Dunn 1997), and 9% based on the Kaufman Brief Intelligence Test–2nd Edition (Kaufman and Kaufman 2004).

Observational ratings of youth behavior. During the final year of data collection, an IRB-approved modification added observational ratings of the youth behavior, using the Guide to the Assessment of Test Session Behavior (GATSB) (Glutting and Oakland 1993). The GATSB is an age-normed observational system where the rater scores 29 behaviors on a 0- to 2-point scale. The GATSB generates three scores: Inattention, Uncooperative Mood, and Avoidant Behavior, as well as a Total Problems score. At the end of the interview day, the KSADS interviewer completed the GATSB to describe the youth's behavior. In addition, the second research assistant who supervised the youth while the caregiver was completing the KSADS interview also filled out a second, independent GATSB.

Analytic plan

Descriptive statistics examined rates of caregiver and youth credibility, and kappa quantified whether youth credibility was associated with caregiver credibility. Correlations measured the extent to which credibility was associated with demographic variables, youth diagnoses, or parent mood symptoms. Multiple regressions using credibility as the dependent variable examined what factors made unique contributions to credibility. Three sets of analyses tested the effect of credibility on the reliability or validity of information from the caregiver or youth by stratifying on credibility: (a) Feldt's (1969) procedure tested whether the internal consistency reliability changed between different levels of credibility; (b) *z*-tests of independent *r* values (Cohen and Cohen 1983) tested potential differences in validity coefficients comparing mood

and behavior checklist scores to diagnoses and interview-based mood ratings; and (c) Hanley and McNeil's (1983) procedure for comparing Receiver Operating Characteristic (ROC) analyses tested whether the diagnostic discriminative validity changed significantly between good, fair, and poor credibility informants. All analyses report uncorrected p values; p values denoted by two or more asterisks signify $p < 0.005$ and would survive even Bonferroni correction for the ''study-wise'' rate of all significance tests run in the course of this investigation.

Results

Description of participants

Table 1 provides information about demographics, youth diagnoses, and study scale descriptives.

Perceived credibility ratings

At the end of the KSADS, interviewers rated 63% of caregivers ''good'' credibility, 31% ''fair,'' and 6% ''poor'' versus 24% of youths ''good,'' 47% ''fair,'' and 30% ''poor.'' Credible youths tended to have credible caregivers: chi-squared (4 degrees of freedom) = 31.69, $p < 0.00005$; but there were still frequent mismatches: kappa = 0.11. Interviewers perceived caregivers as much more credible on average for the young children (ages 5 to 10 years): Cohen's $d = 1.29$, $p < 0.00005$; in contrast, caregivers were only slightly more credible on average for the older youths: $d = 0.28$, $p < 0.0005$.

Correlates of caregiver credibility

Table 1 provides correlations between caregiver credibility and family, youth, and caregiver characteristics. All potential correlates showed distributions with acceptable skew and kurtosis values that fell within the range where Pearson correlation and multiple regression values tend to be robust, and there were no substantive outliers based on standard regression diagnostics (Tabachnick and Fidell 2007). Significant correlates of caregiver credibility included younger youth age, better family functioning (measured as either the interview-rated GFES or the caregiver-rated FAD Total), better youth functioning (based on the interview global assessment of functioning), higher caregiver income or education, fewer children, and lower concerns about youth depression or manic symptoms. Regression analyses indicated that a combination of factors could explain 17% of the variance in caregiver credibility ($p < 0.00005$), with family functioning ($r_{part} = 0.26$), youth age ($r_{part} = -0.21$), credibility of the youth ($r_{part} = 0.18$), and caregiver education ($r_{part} = 0.09$) making significant unique contributions. Caregiver credibility was unrelated to caregiver mood symptoms, youth cognitive ability, or independent observations of youth behavior problems.

Correlates of youth credibility

Table 1 also provides correlations for youth credibility ratings with family, youth, and caregiver characteristics, along with tests of whether the correlations significantly differ between youth versus caregiver credibility. Significant correlations of youth credibility were mostly different from the predictors of caregiver credibility. Older age youth, female youth, not having a diagnosis of ADHD or bipolar disorder, lower CBCL Externalizing or Attention Problems, higher caregiver education, having a male primary caregiver, lower caregiver report of manic symptoms, and higher self-report of manic or depressive symptoms all were significantly associated with greater levels of youth credibility. The correlation with self-reported manic symptoms was small and suggests that the subset of youths with insight into their behavior were perceived as slightly more credible. A subset of cases also completed a brief intelligence test and had observational ratings of behavior available. Greater youth credibility was strongly associated with higher cognitive ability ($r = 0.40$) and less behavior problems during the KSADS interview ($r = -0.25$) or when watched by a different person while the caregiver was completing the KSADS ($r = -0.38$). Regression analyses indicated that factors could account for 22% of the credibility in youth report ($p < 0.0005$), with age being the strongest predictor. Controlling for youth age eliminated all other correlates except for caregiver credibility ($r_{part} = 0.18$), with age remaining a powerful predictor ($r_{part} = 0.39$). For the subset with cognitive ability and behavioral observations available, the regression explained a similar amount of variance, with age, cognitive ability, and observational ratings of behavior, each making unique contributions, but caregiver credibility was no longer significant. Comparing the correlates of youth credibility to those of caregiver credibility found that youth credibility was significantly more linked to youth cognitive ability, youth behavior during the interview, and youth diagnoses of ADHD or mood problems, whereas caregiver credibility was significantly more associated with family functioning or socioeconomic status.

Effect of credibility on reliability

Table 2 presents the internal consistency estimates for the GBI scales reported by the parent and youth. Cronbach's alpha was significantly higher ($p < 0.05$ based on Feldt's test) for the poor credibility caregivers on the depression scale compared with both the fair and good credibility, and there was a similar trend for poor versus good credibility on the hypomanic/biphasic scale. This suggests that poor credibility informants answered with a response set rather than reflecting on the content of each item. Consistent with this possibility, caregiver report on the FAD showed a significant pattern in the opposite direction, with lower internal consistency for the poor credibility caregivers ($p < 0.05$). The FAD includes 10 items that are reverse keyed, so selecting the same response option for all items would lower the reliability estimate for the FAD, whereas it would raise reliability on scales such as the GBI that do not use any reverse keying. Similarly, there was a tendency for the interview-based ratings of the severity of mood symptoms to be more internally consistent when interviewing good credibility rather than poor credibility informants ($p = 0.0903$ on the KMRS and 0.1401 on the KDRS). There were no trends for the association between youth credibility and internal consistency of youth report on the GBI and between youth credibility and interview ratings of mood (all $p > 0.25$).

Effect of credibility on criterion and discriminative validity

Ratings of caregiver credibility were related to the validity of caregiver report on mood and behavior checklists. Criterion validity coefficients for caregiver-reported manic symptoms changed from 0.27 for poor credibility to 0.50 for good credibility when comparing PGBI to KMRS ratings and from 0.52 to 0.64 for PGBI compared with KDRS ratings; however, these did not achieve statistical significance because of the small number of poor credibility caregivers. Caregiver–youth and caregiver–teacher correlations all significantly increased when comparing good credibility to

Table 1. Correlations Between Caregiver and Youth Credibility Ratings with Demographics, Family Characteristics, Clinical Features, and Ratings of Youth Emotions and Behavior

Variable	*M (SD) or percentage*	*Credibility rating*		*t-Test of difference*
		Caregiver	*Youth*	
Youth demographics				
Youth age (years)	10.8 (3.5)	−0.17****	0.43****	−13.29****
Youth female gender	39%	−0.05	0.14***	−3.74***
White Ethnicity	22%	0.09*	0.06	0.71
Intelligent Quotient Score[a]	88.3 (13.8)	0.01	0.40****	−3.63***
Family characteristics and functioning				
Number of siblings	3.1 (1.8)	−0.11**	−0.02	−1.84
Reporter is biological mother	78%	−0.05	−0.05	0.09
Reporter is biological father	5%	−0.04	0.06	−2.02*
Has a male primary caregiver	6%	−0.05	0.08*	−2.61*
Caregiver with youth most or all of past year?	97%	0.04	−0.02	1.12
Estimated income	3.3 (2.2)	0.14**	0.04	1.82
Caregiver education	4.3 (1.2)	0.14**	0.11*	0.61
Occupational status	2.4 (2.6)	0.14**	0.05	1.68
FAD Total Problems (caregiver rated)	2.0 (0.4)	−0.16****	−0.02	−2.84**
Global Family Environment (interviewer rated)	67.4 (12.0)	0.28****	0.02	5.34****
Caregiver Beck Depression Inventory (re: self)	9.1 (8.3)	−0.02	0.03	−1.02
Caregiver Mood Disorder Questionnaire (re: self)	3.2 (3.1)	−0.06	−0.01	−0.95
Youth clinical features				
Attention-deficit/hyperactivity disorder	64%	0.02	−0.18****	4.05***
Oppositional defiant disorder	39%	0.02	0.02	0.00
Unipolar depression or dysthymia	28%	−0.06	0.12**	−3.52***
Any anxiety disorder	27%	−0.02	0.16***	−3.53***
Bipolar spectrum diagnosis	19%	0.05	−0.09*	2.76*
Conduct disorder	14%	−0.14**	−0.06	−1.57
Posttraumatic stress disorder	8%	−0.08	0.15***	−4.54****
Prior clinical diagnosis of ADHD	34%	0.10*	−0.21****	6.31****
Number of Axis I Diagnoses	2.7 (1.4)	−0.07	−0.01	−1.17
Current Global Assessment of Functioning	52.1 (8.1)	0.13**	0.08*	0.87
Physical abuse	17%	−0.05	0.03	−1.48
Sexual abuse	19%	0.03	0.07	−0.73
Neglect	16%	0.02	−0.01	0.51
Caregiver-reported emotion and behavior				
Child Behavior Checklist Externalizing	69.7 (9.6)	−0.07	−0.19****	2.33*
Child Behavior Checklist Internalizing	63.4 (10.2)	−0.04	0.00	−0.77
Child Behavior Checklist Attention Problems	69.0 (11.6)	−0.02	−0.16***	2.54*
PGBI Hypomanic/Biphasic	20.1 (14.2)	−0.09*	−0.10*	0.29
PGBI Depression	26.0 (21.6)	−0.09*	0.05	−2.86**
Youth-reported emotion and behavior (ages 11+)				
AGBI Hypomanic/Biphasic	25.1 (15.4)	−0.03	0.12*	−2.14*
AGBI Depression	41.2 (28.7)	−0.07	0.18***	−3.71***
Youth Self-Report Externalizing	59.0 (11.7)	−0.06	0.05	−1.65
Youth Self-Report Internalizing	56.9 (12.7)	−0.06	0.12*	−2.57*
Youth Self-Report Attention Problems	60.4 (10.6)	−0.02	0.06	−1.08
Interviewer-Rated Mood Symptoms				
KSADS Mania Rating Scale Total	20.5 (9.7)	0.04	−0.06	1.93
KSADS Depression Rating Scale Total	21.9 (8.9)	−0.04	0.18****	−4.25****
Guide to the Assessment of Test Session Behavior-First Observer (KSADS Interviewer)[b]				
Total problems	81.4 (14.0)	−0.06	−0.25*	1.63
Inattentive	58.3 (13.9)	−0.02	−0.40****	3.26**
Avoidant	59.5 (12.1)	0.06	−0.39****	3.86***
Uncooperative	55.7 (10.7)	−0.12	−0.26*	1.11
GATSB-Second Observer[b]				
Total problems	73.7 (17.8)	−0.08	−0.38****	2.53*
Inattentive	54.1 (13.4)	−0.01	−0.33***	2.72*
Avoidant	61.5 (14.0)	−0.09	−0.27**	1.47
Uncooperative	56.1 (14.4)	−0.08	−0.22*	1.16
Teacher Report Form[c]				
Externalizing	64.7 (10.5)	0.01	−0.03	0.45
Internalizing	57.7 (10.3)	0.08	0.05	0.29
Attention problems	63.6 (9.5)	0.08	0.03	0.62

[a]Intelligent Quotient score = full scale score, $n = 116$.

[b]Guide to the Assessment of Test Session Behavior, $n = 109$.

[c]Teacher Report Form, $n = 249$.

*$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$, ****$p < 0.00005$, two tailed.

M = mean; SD = standard deviation; FAD = family assessment device; PGBI = Parent General Behavior Inventory; AGBI = Adolescent General Behavior Inventory; KSADS = Kiddie Schedule for Affective Disorders and Schizophrenia; ADHD = attention-deficit/hyperactivity disorder.

Table 2. Internal Consistency Reliability Estimates as a Function of the Perceived Credibility of the Caregiver (n = 646)

Variable	Credibility			Feldt's test
	Good	*Fair*	*Poor*	
Caregiver-reported emotion and behavior				
PGBI Hypomanic/Biphasic	0.921	0.927	0.943	Poor > good, $p = 0.0886$
PGBI Depression	0.950	0.959	0.974	Poor > good, $p = 0.0032$; poor > fair, $p = 0.0378$
FAD Total (includes 10 reverse keyed items)	0.915	0.900	0.875	Good > poor, $p = 0.0485$
Interviewer-Rated Mood Symptoms				
KSADS Mania Rating Scale	0.932	0.919	0.907	Good > poor, $p = 0.0903$
KSADS Depression Rating Scale	0.872	0.853	0.835	No trends

poor credibility caregivers, and this pattern was found across ratings of externalizing, internalizing, and attention problems as well as for manic and depressive symptoms (all $p < 0.05$) (cf. Achenbach et al. 1987). With regard to discriminative validity, areas under the curve in ROC analyses change from 0.63 (nonsignificant) to 0.81 ($p < 0.0005$) for poor versus good credibility when comparing PGBI scores to bipolar diagnoses.

Similar patterns were observed with youth credibility, with the criterion correlations for the GBI depression score rising from 0.27 to 0.49, and the hypomanic/biphasic score from 0.39 to 0.43 when comparing poor versus good credibility youths, with the ROCs against bipolar diagnoses being significant for the good but not poor credibility youths. These patterns were not due to changes in the internal consistency of checklist scores, as internal consistency either stayed the same or increased for the poor credibility informants.

Discussion

At the end of a day-long interview, clinical interviewers judged caregivers to be credible informants more often than youths. The correlates of caregiver reliability and youth reliability each had face validity, but there was little overlap in the predictors. Caregivers were perceived as most credible when families were functioning better (consistent with Hawley and Weisz 2003), when they were reporting about younger children, when the youth was also perceived as credible, and when the caregiver was more educated. Caregiver mood symptoms, being the biological mother, and various other plausible correlates showed no significant association with caregiver credibility. The lack of relations between credibility and caregiver mood stands in contrast to prior findings that the validity of caregiver report might be attenuated by caregiver stress, although these effects have tended to be small (Richters 1992; Youngstrom et al. 2000). Perceived youth credibility markedly increased with youth age, with cognitive ability and independent observations of youth behavior problems explaining additional variance. If the youth met criteria for ADHD or a bipolar diagnosis, then their perceived credibility tended to be significantly lower.

Regression models using demographic and clinical features explained moderate amounts of variance with high degrees of statistical significance for both caregiver and youth credibility. These plausible and often substantial associations corroborate the validity of clinical judgments about credibility. However, the predictions fell far short of what would be needed to classify informants based on these variables instead of using clinical judgment to rate credibility directly. Similarly, the changes in validity of caregiver or youth-reported ratings, although often statistically significant, were never so large as to justify substantial changes in the interpretation of information, such as discounting or ignoring an informant entirely. Informants with poor credibility still usually provide ratings with some validity, albeit moderately less valid than corresponding reports from informants with good credibility. Ratings of manic symptoms in the youth varied from moderate to high validity, for example.

On the other hand, clinical interviewers appear to be able to integrate multiple pieces of information gathered during a semi-structured interview to arrive at a valid decision about whether informants have good, fair, or poor credibility. Informants judged to have ''good'' credibility showed validity coefficients equal to or higher than the benchmarks reported in the literature for diagnostic efficiency and cross-informant agreement. Conversely, informants judged to have ''poor'' credibility demonstrated validity coefficients that were sometimes significantly lower than the credible informants in the same setting, as well as below published benchmarks. These relationships were significant despite safeguards such as blinded ratings by independent raters, and they provide strong evidence for the validity of clinical judgment about the credibility of informants.

Decreased reliability is a frequent cause of decreased validity in assessment, but in this case the reports from informants with poor credibility actually had the same or higher internal consistency. This suggests that informants were forming a response set and describing behavior problems in a global, uniform way, thus yielding highly consistent but less valid reports.

Limitations and future directions

Investigation of credibility was based on a secondary analysis of data originally gathered for other purposes. Although sample size remained considerable, a variety of measures that were helpful in examining correlates of credibility were only available on subsets of cases. There also are many factors that may have a large effect on rater credibility, but these were not directly measured in this study. Candidates include constructs such as social desirability, denial, malingering, and other factors that can systematically influence scores (Guion 1998).

The actual rating of credibility was also simplistic, asking the interviewer to make a global evaluation with only a few options. Many raters opted to use decimals to convey additional gradations of credibility, indicating that more nuanced perceptions could be quantified. It is also likely that more elaborated rating systems could focus attention on different aspects of credibility, providing

more detail about facets such as demoralization, malingering, impression management, or lack of insight. Moving from global, unstructured impressions to more objectified, semistructured ratings often achieves enhanced reliability and validity (Anastasi 1988). This would be a promising area for future development of rating scales, as even simple global ratings demonstrated moderate criterion correlations and changes in the validity of information provided by caregivers and youths.

More work is needed to refine assessment approaches to consider the credibility of the informant, rather than always basing algorithms on the ''average'' caregiver or ''average'' youth. It remains to be determined whether clinical decision making would be better enhanced by determining when to discard information, because a particular informant seems to have poor credibility, versus keeping the information but adjusting the weight or interpretation. Future work should also investigate whether the effects of credibility are equally powerful across different diagnoses, different domains of functioning, and different rating scales. Present results suggest that the effects of caregiver versus youth credibility may not be the same for externalizing versus internalizing or attention problems, and similar issues may apply to conditions such as anxiety disorders (where self-report may often be highly accurate) (Frick et al. 1994) versus conduct disorder (with a high risk of denial in self-report) or psychosis (where insight may rapidly be compromised) (Pini et al. 2001; Youngstrom et al. 2004a). Similarly, some rating scales may be more susceptible to the effects of changing informant credibility, because of issues such as rater burden and reading level as well as perceived social desirability (Garb 1998). Differences in credibility are also likely to be associated with patterns of agreement between informants about the youths' functioning (De Los Reyes et al., 2011). The majority of participants came from low-income families and impoverished school systems, as reflected in the distribution of cognitive ability scores. It would be helpful to investigate correlates of credibility in samples with different demographics to understand the extent to which SES might moderate credibility and its associated features.

Conclusion

Overall, present findings indicate that not all caregivers, and all youths, are seen as equally credible by clinicians, nor should they be. Clinical judgments about credibility showed plausible relationships with youth, caregiver, and family characteristics as well as measurable changes to the reliability and validity of information they provided on standard checklists. Clinicians appear to be able to gauge when information from a particular source may be suspected. The next wave of research should refine how to quantify judgments of credibility and develop evidence-based approaches for integrating these data into the assessment and decision-making process.

Clinical significance

No prior research has investigated whether clinical judgments about the credibility of adults or youths are related to the reliability or validity of the information they provide. Findings add to knowledge about factors associated with credibility and document the extent to which poor credibility is linked with reduced criterion validity. Clinicians will encounter caregivers or youths with compromised credibility because of various circumstances, and global clinical judgments about credibility are linked with changes in response set, shifts in degree of cross-informant agreement, and changes in the diagnostic validity of information received. However, even poor credibility did not totally invalidate the youth or caregiver report on any instrument. Clinicians should keep this in mind when integrating information from various informants. Completely discounting or ignoring information from a person judged to have poor credibility will overcorrect and often result in less accurate decisions. Until more precise and generalizable algorithms are developed, a reasonable clinical strategy would be to pay attention to credibility, but to make more fine-grained adjustments in interpretation rather than dropping a set of scores entirely when credibility is assessed as being poor. When credibility is only ''fair,'' reports may need to be taken with some degree of circumspection, and even ''poor'' credibility informants provided information that remained statistically valid.

Acknowledgments

The authors thank the families who participated in this research. This work was supported in part by NIH R01 MH066647 (PI: E. Youngstrom).

Disclosures

E. Youngstrom has received travel support from Bristol-Myers Squibb. Dr. Findling receives or has received research support, acted as a consultant and/or served on a speaker's bureau for Abbott, Addrenex, AstraZeneca, Biovail, Bristol-Myers Squibb, Forest, GlaxoSmithKline, Johnson & Johnson, KemPharm Lilly, Lundbeck, Neuropharm, Novartis, Noven, Organon, Otsuka, Pfizer, Rhodes Pharmaceuticals, Sanofi-Aventis, Schering-Plough, Seaside Therapeutics, Sepracore, Shire, Solvay, Sunovion, Supernus Pharmaceuticals, Validus, and Wyeth. The other authors have no financial interests to disclose.

References

Achenbach TM. Empirically based assessment and taxonomy: Applications to clinical research. Psychol Assess 7:261–274, 1995.

Achenbach TM, McConaughy SH, Howell CT: Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. Psychol Bull 101:213–232, 1987.

Achenbach TM, Rescorla LA: Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: University of Vermont; 2001.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th ed., Text Revision (DSM-IV-TR). Washington, DC: American Psychiatric Association; 2000.

Anastasi A: Psychological Testing. New York: MacMillan; 1988.

Arkes HR: Costs and benefits of judgment errors: Implications for debiasing. Psychol Bull 110:486–498, 1991.

Axelson DA, Birmaher BJ, Brent D, Wassick S, Hoover C, Bridge J, Ryan N: A preliminary study of the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children Mania Rating Scale for children and adolescents. J Child Adolesc Psychopharmacol 13:463–470, 2003.

Barkley RA: Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment, 2nd edition. New York: Guilford; 1998.

Beck AT, Steer RA: Beck Depression Inventory Manual. San Antonio, TX: The Psychological Corporation; 1987.

Bird HR, Gould MS, Staghezza B: Aggregating data from multiple informants in child psychiatry epidemiological research. J Am Acad Child Adolesc Psychiatry 31:78–85, 1992.

Byles J, Byrne C, Boyle MH, Offord DR: Ontario child health study: Reliability and Validity of the General Functioning Subscale of the Mcmaster Family Assessment device. Fam Process 27:97–104, 1988.

Carlson GA, Blader J. Diagnostic implications of informant disagreement for manic symptoms. J Child Adolesc Psychopharmacol 21:399–405, 2011.

Carlson GA, Youngstrom EA: Clinical implications of pervasive manic symptoms in children. Biol Psychiatry 53:1050–1058, 2003.

Chavez LM, Shrout PE, Alegria M, Lapatin S, Canino G: Ethnic differences in perceived impairment and need for care. J Abnorm Child Psychol 38:1165–1177, 2010.

Cohen J, Cohen P: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition. Hillsdale, NJ: Lawrence Erlbaum; 1983.

Danielson CK, Youngstrom EA, Findling RL, Calabrese JR: Discriminative validity of the general behavior inventory using youth report. J Abnorm Child Psychol 31:29–39, 2003.

Dawes RM, Faust D, Meehl PE: Clinical versus actuarial judgment. Science 243:1668–1674, 1989.

De Los Reyes A: Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. J Clin Child Adolesc Psychol 40:1–9, 2011.

De Los Reyes A, Henry DB, Tolan PH, Wakschlag LS: Linking informant discrepancies to observed variations in young children's disruptive behavior. J Abnorm Child Psychol 37:637–652, 2009.

De Los Reyes A, Kazdin AE: Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. Psychol Bull 131:483–509, 2005.

De Los Reyes A, Youngstrom EA, Swan AJ, Youngstrom JK, Feeny NC, Findling RL: Informant discrepancies in clinical reports of youths and interviewers' impressions of the reliability of information. J Child Adolesc Psychopharmacol 21:417–424, 2011.

Dunn LM, Dunn LM: Examiner's Manual for the Peabody Picture Vocabulary Test, 3rd Edition. Circle Pines, MN: American Guidance Service; 1997.

DuPaul GJ, Power TJ, McGoey KE, Ikeda MJ, Anastopoulos AD: Reliability and validity of parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms. J Psychoeduc Assess 16: 55–68, 1998.

Epkins CC: Teachers' ratings of inpatient children's depression, anxiety, and aggression: A preliminary comparison between inpatient-facility and community-based teachers' ratings and their correspondence with children's self-reports. J Clin Child Psychol 24:63–70, 1995.

Feldt LS: A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika 34:363–373, 1969.

Freeman AJ, Youngstrom EA, Freeman MJ, Youngstrom JK, Findling RL. Is caregiver-adolescent disagreement due to differences in thresholds for reporting manic symptoms? J Child Adolesc Psychopharmacol 21:425–432, 2011.

Frick PJ, Silverthorn P, Evans C: Assessment of childhood anxiety using structured interviews: Patterns of agreement among informants and association with maternal anxiety. Psychol Assess 6:372–379, 1994.

Garb HN: Race bias, social class bias, and gender bias in clinical judgment. Clin Psychol: Sci Pract 4:99–120, 1997.

Garb HN: Studying the Clinician: Judgment Research and Psychological Assessment. Washington, DC: American Psychological Association; 1998.

Geller B, Zimerman B, Williams M, Bolhofner K, Craney JL, DelBello MP, Soutullo C: Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. J Am Acad Child Adolesc Psychiatry 40:450–455, 2001.

Glutting J, Oakland T: Guide to the Assessment of Test Session Behavior for the WISC-III and WIAT. San Antonio: The Psychological Corporation; 1993.

Gonzalez JM, Alegria M, Prihoda TJ, Copeland LA, Zeber JE: How the relationship of attitudes toward mental health treatment and service use differs by age, gender, ethnicity/race and education. Soc Psychiatry Psychiatr Epidemiol 46:45–57, 2011.

Groth-Marnat G: Handbook of Psychological Assessment, 3rd edition. New York: Wiley; 1999.

Guion RM: Assessment, measurement, and prediction for personnel decisions. Hillsdale, NJ: Erlbaum; 1998.

Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843, 1983.

Hartley AG, Zakriski AL, Wright JC: Probing the depths of informant discrepancies: contextual influences on divergence and convergence. J Clin Child Adolesc Psychol 40:54–66, 2011.

Hawley KM, Weisz JR: Child, parent, and therapist (dis)agreement on target problems in outpatient therapy: the therapist's dilemma and its implications. J Consult Clin Psychol 71:62–70, 2003.

Hazell PL, Lewin TJ, Carr VJ: Confirmation that child behavior checklist clinical scales discriminate juvenile mania from attention deficit hyperactivity disorder. J Paediatr Child Health 35:199–203, 1999.

Henry GK, Heilbronner RL, Mittenberg W, Enders C and Domboski K: Comparison of the MMPI-2 restructured demoralization scale, depression scale, and malingered mood disorder scale in identifying non-credible symptom reporting in personal injury litigants and disability claimants. Clin Neuropsychologist 23:153–166, 2009.

Hirschfeld RM, Williams JBW, Spitzer RL, Calabrese JR, Flynn L, Keck PEJ, Lewis L, McElroy SL, Post RM, Rapport DJ, Russell JM, Sachs GS, Zajecka J: Development and validation of a screening instrument for bipolar spectrum disorder: The mood disorder questionnaire. Am J Psychiatry 157:1873–1875, 2000.

Hudziak JJ, Derks EM, Althoff RR, Copeland W, Boomsma DI: The genetic and environmental contributions to oppositional defiant behavior: A multi-informant twin study. J Am Acad Child Adolesc Psychiatry 44:907–914, 2005.

Jensen PS, Rubio-Stipec M, Canino G, Bird HR, Dulcan MK, Schwab-Stone ME, Lahey BB: Parent and child contributions to diagnosis of mental disorder: Are both informants always necessary? J Am Acad Child Adolesc Psychiatry 38:1569–1579, 1999.

Kaufman AS, Kaufman NL: Kaufman Brief Intelligence Test, 2nd edition. San Antonio, TX: Pearson, 2004.

Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, Williamson D, Ryan N: Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): Initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 36:980–988, 1997.

Kraemer HC: Evaluating Medical Tests: Objective and Quantitative Guidelines. Newbury Park, CA: Sage Publications; 1992.

Kraemer HC, Lowe KK, Kupfer DJ: To Your Health: How to Understand What Research Tells Us About Risk. New York: Oxford University Press; 2005.

Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ: A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. Am J Psychiatry 160:1566–1577, 2003.

Loeber R, Green SM, Lahey BB, Stouthamer Loeber M: Optimal informants on childhood disruptive behaviors. Dev Psychopathol 1:317–337, 1989.

Mash EJ, Hunsley J: Evidence-based assessment of child and adolescent disorders: issues and challenges. J Clin Child Adolesc Psychol 34:362–379, 2005.

Meehl PE: Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. Minneapolis: University of Minnesota Press; 1954.

Phares V: Where's poppa? The relative lack of attention to the role of fathers in child and adolescent psychopathology. Am Psychol 47:656–664, 1992.

Phares V: Accuracy of informants: Do parents think that mother knows best? J Abnorm Child Psychol 25:165–171, 1997.

Pini S, Dell'Osso L, Amador XF: Insight into illness in schizophrenia, schizoaffective disorder, and mood disorders with psychotic features. Am J Psychiatry 158:122–125, 2001.

Rey JM, Singh M, Hung SF, Dossetor DR, Newman L, Plapp JM, Bird KD: A global scale of measure the quality of the family environment. Arch Gen Psychiatry 54:817–822, 1997.

Reynolds CR, Kamphaus R: BASC-2 Behavior Assessment System for Children. Circle Pines, MN: American Guidance Service; 2004.

Richters JE: Depressed mothers as informants about their children: A critical review of the evidence for distortion. Psychol Bull 112:485–499, 1992.

Sellbom M, Ben-Porath YS: Mapping the MMPI-2 Restructured Clinical Scales onto normal personality traits: Evidence of construct validity. J Pers Assess 85:179–187, 2005.

Sherman RA, Nave CS, Funder DC: Situational similarity and personality predict behavioral consistency. J Pers Soc Psychol 99: 330–343, 2010.

Spitzer RL: Psychiatric diagnosis: Are clinicians still necessary? Compr Psychiatry 24:399–411, 1983.

Tabachnick BG, Fidell LS: Using Multivariate Statistics, 5th edition. Boston: Allyn and Bacon; 2007.

The Psychological Corporation. Wechsler Abbreviated Scale of Intelligence Manual. San Antonio: Harcourt Brace and Company; 1999.

Youngstrom EA, Findling RL, Calabrese JR: Who are the comorbid adolescents? Agreement between psychiatric diagnosis, parent, teacher, and youth report. J Abnorm Child Psychol 31:231–245, 2003.

Youngstrom EA, Findling RL, Calabrese JR: Effects of adolescent manic symptoms on agreement between youth, parent, and teacher ratings of behavior problems. J Affect Disord 82:S5–S16, 2004a.

Youngstrom EA, Findling RL, Calabrese JR, Gracious BL, Demeter C, DelPorto Bedoya D, Price M: Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. J Am Acad Child Adolesc Psychiatry 43:847–858, 2004b.

Youngstrom EA, Findling RL, Danielson CK, Calabrese JR: Discriminative validity of parent report of hypomanic and depressive symptoms on the general behavior inventory. Psychol Assess 13:267–276, 2001.

Youngstrom EA, Joseph MF, Greene J: Comparing the psychometric properties of multiple teacher report instruments as predictors of bipolar disorder in children and adolescents. J Clin Psychol 64: 382–401, 2008.

Youngstrom EA, Loeber R, Stouthamer-Loeber M: Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. J Consult Clin Psychol 68:1038–1050, 2000.

Address correspondence to:
Eric A. Youngstrom, Ph.D.
Department of Psychology
University of North Carolina
CB #3270, Davie Hall
Chapel Hill, NC 27599-3270

E-mail: eay@unc.edu