A New Approach to Integrating Data From Multiple Informants in Psychiatric Assessment and Research: Mixing and Matching Contexts and Perspectives

Helena C. Kraemer, Ph.D.

Jeffrey R. Measelle, Ph.D.

Jennifer C. Ablow, Ph.D.

Marilyn J. Essex, Ph.D.

W. Thomas Boyce, M.D.

David J. Kupfer, M.D.

Objective: When there exists no single source of information (informant) to validly measure a characteristic, it is typically recommended that data from multiple informants be used. In psychiatric assessment and research, however, multiple informants often provide discordant data, which further confuse the measurement. Strategies such as arbitrarily choosing one informant or using the data from all informants separately generate further problems. This report proposes a theory to explain observed patterns of interinformant discordance and suggests a new approach to using data from multiple informants to measure characteristics of interest.

Method: Using the example of assessment of developmental psychopathology in children, the authors propose a model in which the choice of informants is based on conceptualizing the contexts and perspectives that influence expression of the characteristic of interest and then identifying informants who represent those contexts and perspectives in such a way as to have the weaknesses of one informant canceled by the strengths of another.

Results: Applications of this approach to several datasets indicate that when these principles are followed, a more reliable and valid consensus measure is obtained, and failure to obtain a reliable, valid measure is indicative of some deviation from the principles.

Conclusions: In obtaining a consensus measure, the issue is not determining how many informants are needed but choosing the right set of informants.

(Am J Psychiatry 2003; 160:1566–1577)

In many clinical and research contexts, certain characteristics of patients that are often vital to understanding development or psychopathology are difficult to measure well. Such a situation pertains when data on the characteristic in question can be obtained only by asking subjects or informants rather than by using more objective measures such as those based on direct expert observation or on laboratory measurements from imaging or tissue sampling. This difficulty is particularly problematic, for example, in working with subjects who have impaired cognition or communicative abilities (e.g., patients with Alzheimer's disease), when any single individual has insufficient access to the necessary information (e.g., in assessing family history of a disorder for genetic studies), or when the subjects are unwilling to impart the necessary information or unable to provide reliable data (e.g., in assessment of compliance with treatment or of illicit drug use). A particularly salient situation is the assessment of young children, where parents, teachers, and clinicians are often the sources of information about the child. In light of federal guidelines mandating adequate consideration of children in formulating research studies (1), such issues accrue an even greater importance.

The purpose of this report is to propose a theory explaining observed patterns of interinformant discordance and, by doing so, to indicate a more valid approach to measuring the characteristics of interest. While the approach described in this article is generally applicable, we will illustrate both the scope of the problem and the application of the proposed methods by focusing on the assessment of developmental psychopathology in children. We apply this approach to data from three studies, demonstrating when and how it works and illustrating results when it succeeds and when it fails. The focus here is on measuring a characteristic of the subject, not an informant's perception of the characteristic, for in that case the informant's report would be employed without hesitation. Thus, if one is interested in the health of the child, the methods we describe might usefully be considered, but if the mother's perception of the child's health is what matters, the mother's report would be used whether or not it corresponded precisely to the actual health of the child.

Background: Child Risk and Psychopathology

For the assessment of childhood symptoms and psychopathology, there is no gold standard measure. By a gold standard measure, we mean a measurement procedure for which the accuracy (validity) and precision (reli-

ability) are sufficiently high that the measure's utility in clinical decision making or research applications is indisputable (2). In the absence of such a gold standard measure—as in the assessment of child health, temperament, impairment, or socioemotional behavior—there is broad agreement that assessment requires data from multiple informants, e.g., the perspectives of the children themselves; their parents, teachers, and clinicians; and other individuals who know the child well (3–5).

When multiple informants provide information, however, a convergence of the data is almost never achieved. Scientists and clinicians must contend with levels of agreement about childhood impairment and psychopathology that are rarely more than moderate in strength. Correlational estimates of cross-informant agreement tend to hover around 0.13 when parents and teachers report on internalizing symptoms and 0.32 when they report on externalizing behavior problems (6–8). The correlation between children's and adults' reports tends not to exceed 0.20 for either form of behavioral difficulty (6, 9). In a study of measures of relationships and experiences in twins (10), the interinformant correlations ranged from 0.14 to 0.63, with most in the 0.30–0.40 range.

Faced with low levels of interinformant agreement, researchers and clinicians must reconcile discrepant reports, yet there is little or no consensus regarding appropriate means to effect such reconciliation (11). Little is known about the sources of the discrepancies, although it is likely that low levels of cross-informant agreement imply that childhood functioning and impairment can be understood best as representing the separate and combined influences of 1) children's actual characteristics (e.g., traits, symptoms, competencies), 2) the context (or situations) in which children are observed, 3) the perspectives (or biases) of the different informants, and 4) error of measurement (12, 13). The work of Achenbach (14), Kazdin (8), Offord and his colleagues (11, 15), and numerous others (3, 7, 16) has highlighted the importance of understanding the basis of informant agreement and disagreement as it pertains to the assessment of childhood dysfunction. As a body of work, this research draws attention to the fact that the reports of different informants yield markedly varied prevalence rates for most major childhood psychopathology (17, 18). Moreover, when treatment decisions are predicated on informants' reports, a majority of children with demonstrated service needs go untreated (11). Advances in understanding and treating childhood psychopathology have thus been impeded by the lack of an adequate solution to this core problem. Insufficient solutions include the selection of a sole or "optimal" informant, the concurrent use of data from all informants, and the aggregation of multi-informant data.

The Optimal Informant

In both clinical and research applications, one informant may be selected as optimal, yet the criteria for choosing the optimal informant for each characteristic are not clear. To treat one source of information arbitrarily as the ideal informant increases the risk of obtaining the right answer to the wrong question. For example, the Infant Health and Development Program (19), a multisite, randomized clinical trial of a behavioral intervention to reduce health risks associated with low birth weight and prematurity, assessed children's health status on the basis of retrospective maternal reports of illnesses. Such reports proved, however, to be highly correlated with the mothers' vocabulary, education, race, and age. On the basis of the maternal reports, it appeared that lower family socioeconomic status was associated with better child health (19), a relationship that most investigators found untenable. Consequently, when it was shown that the reported health of the child was significantly poorer in the treatment group than in the control group, it was not clear whether the effect of the treatment was to improve the mother's knowledge of and sensitivity to her child's health (since one component of the treatment was parent education) or to impair the health of the child. If the treatment actually impaired the health of the child, one would question the clinical utility of the treatment, but if it improved the mother's understanding of and sensitivity to her child's health, the treatment would be encouraged.

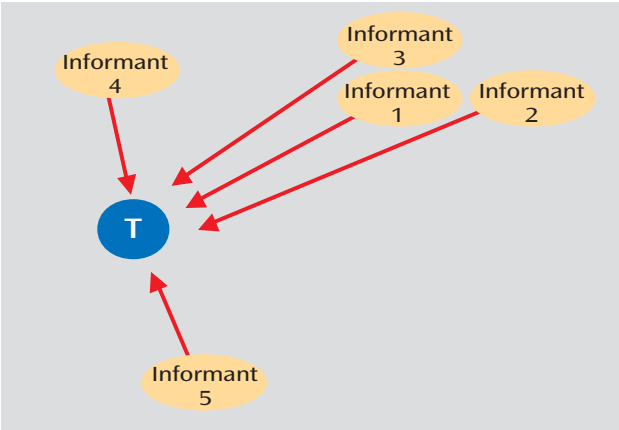
Using Data From All Informants Separately and Simultaneously

In clinical applications, a clinician's decision about whether to treat or how to treat remains unclear when multiple informants disagree on treatment indications. In research applications, on the other hand, a common strategy is to use data from all informants, however discrepant, separately and simultaneously. Thus, in the Multimodal Treatment of Attention Deficit Hyperactivity Disorder Study, an eight-site clinical trial of the effectiveness of a treatment for children with attention deficit hyperactivity disorder (ADHD), the various parent and teacher reports were separately reported as outcomes (20). With no adjustment for multiple testing, this strategy increases the risk of type I errors, i.e., greater likelihood of false positive results, because a separate risk of a false positive result exists for each outcome. With adjustment, the risk of type II error increases the likelihood of false negative results. When the conclusions of a randomized, controlled trial are incongruous because they are based on different measures of the same child characteristic, the trial's message to clinicians and policy makers is ambiguous, undermining the pragmatic value of the trial.

Aggregation

Finally, in clinical and research applications, one might logically consider aggregation strategies, but, again, the selection of such strategies is often arbitrary. Recognizing the problem of multiple outcomes in the Multimodal Treatment of Attention Deficit Hyperactivity Disorder Study discussed in the previous section (20), the investiga-

FIGURE 1. Schematic Representation of Approach to Integrating Data From Multiple Informants in Psychiatric Assessment and Research[a]



[a] T represents the trait gold standard as if it were located in two-dimensional space. The various informants surrounding T can each reliably indicate the direction in which T lies, but none can individually pinpoint the exact location of T. The reports of informants whose perspectives are orthogonal to each other can be used to triangulate the position of T, and the triangulated measure would be the gold standard measure, for it locates T both accurately and at least as precisely as the separate reports of the multiple informants.

tors made two post hoc attempts to aggregate across measures to adjust for the effects of multiple informants. One attempt was based on clinical judgment of whether symptoms reported by all informants were sufficiently few to consider the outcome a clinical success (21). The other was based on psychometric evaluation of the outcomes and was accomplished by entering the reports of multiple informants into a factor analysis, finding one factor related to mothers' reports and another related to teachers' reports, and averaging the effects of the two factors (22). Either aggregation strategy, had it been used a priori, would have led to clearer, more powerful, and less ambiguous results but would not necessarily have led to the same inferences from a clinical or policy view.

As of 2001, approximately 70 studies examining multi-informant variability in the assessment of child psychopathology had been published, yet there remains a paucity of theory describing processes or mechanisms that are capable of explaining and resolving multi-informant discrepancies. Most published empirical results have been descriptive (e.g., levels of variability by type of symptom, correlates of variability), leaving unresolved the major question about the processes required to obtain a more valid measure of the child's characteristic.

A New Theoretical and Practical Approach

Conceptual Model

Our goal was to develop a general procedure for derivation of a measure based on the reports of multiple informers that was closer to a population gold standard measure than are any of the conventional multiple individual informant strategies.

Figure 1 provides a graphic depiction of the concept underlying this general procedure, one very similar to the principles of the Global Positioning System. Suppose that the gold standard (T) resides at some location in a two-dimensional field and that the task is to locate the gold standard as accurately as possible. A gold standard measure is one that would point directly at the location of the gold standard, albeit perhaps with some imprecision (or unreliability), but no gold standard measure is available. Instead, multiple informants are available, as represented by the positions on the perimeter of the field (analogous to the satellites in various orbital positions), each of which can point with relative precision (i.e., reliably) in the general direction of the gold standard, but none of which can locate it accurately (i.e., validly). If one elected to use only the data from informant 1, the search for the gold standard would be narrowed substantially to a single line in the field, but not resolved. If one then elected to add the data from another informant, the least advantageous choice would be informant 2 or 3, or any other informant that shares the same limitations as informant 1. That is, one would not choose another informant whose perspective was "collinear" (or highly correlated) with that of informant 1. Instead, one would choose a second informant whose data could in some manner correct the deficiencies of informant 1's data. In this case, the second informant would be informant 4 or 5, whose perspective is orthogonal to (or independent of) that of informant 1. In a two-dimensional space (east-west and north-south), one would minimally need two reliable informants for triangulation to find T, and the triangulated position would be a gold standard measure because it locates T both accurately and at least as precisely as the data from the multiple informants could separately.

Similarly, if the gold standard were located in a three-dimensional space (east-west, north-south, distance above ground level), one would minimally require three reliable informants to triangulate the position of the gold standard. Once again, optimal selection of the informants would avoid collinearity and maximize orthogonality. As the dimensionality of the problem increases, so also does the minimal number of informants, but the principle remains the same: it is not only the *number* of informants that matters; it is also *how informants are selected*. Moreover the lack of correlation (orthogonality) between informants, to date considered problematic, becomes precisely the phenomenon that facilitates a more valid measure. An infinite number of collinear informants cannot solve the problem, but in an M-dimensional space, M carefully selected informants may. The first challenge is to attempt a definition of dimensionality that clarifies what the dimensions are in the particular problem and hence what "collinear," "orthogonal," and "independent" mean in the se-

lection of multiple informants. We consider this issue in the next section.

Mathematical Model

The preceding discussion, as well as past research, has led to the following mathematical representation. We propose that the response of an informant about the subject's characteristic over a relevant span of time (*I*) is represented as follows:

$$I=T+C+P+E$$

where *T*, *C*, and *P* are three orthogonal dimensions and *E* is random error. In this equation, *I* is the informant's report.

T is the trait dimension, i.e., the characteristic that differs from subject to subject in the population of interest but for each subject is constant over the designated span of time (perhaps the subject's true mean over that span of time). The label "trait" for the characteristic of interest should not be confounded with the general usage of "trait" as an enduring or unchangeable characteristic of the subject. In this model, the designated span of time can be as short or long as is required for the issue under investigation—for example, in asking informants about a subject's symptoms, the span of time may be 2 weeks, 6 months, or 3 years, depending on the focus of the research (e.g., if the study seeks to explore whether the duration of symptoms meets diagnostic criteria). Thus T, the "trait" of interest, may actually be a transient "state" rather than an enduring and unchangeable characteristic of the individual.

C is the context dimension, i.e., the factors related to place and circumstance that influence the subject's expression of T and contribute to test-retest unreliability.

P is the perspective dimension, i.e., the characteristics of the informant that influence his or her assessment of the trait at any point in time and can constitute the major source of interobserver unreliability.

E is the error of measurement, i.e., the effects of random factors that are not related to subject, context, or perspective but that are perhaps related to the method by which the information is obtained. These effects constitute a major contribution to intraobserver unreliability.

Drawing on fundamental psychometric principles, we know that:

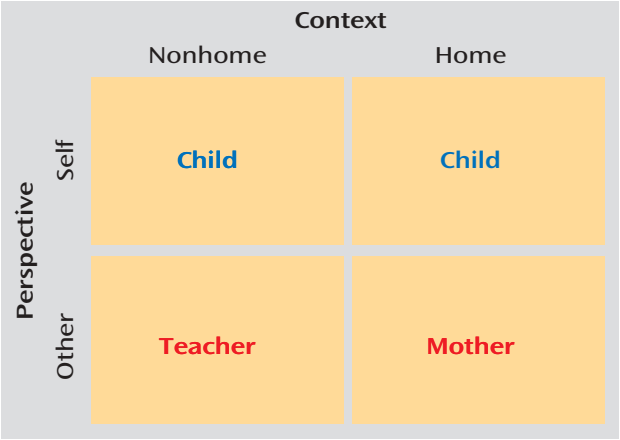
1. The classic definition of the reliability of the informant in a specific population of subjects is the percentage of total variance of I that comes from T+C+P.
2. The classic definition of the validity of I for T is the percentage of the total variance of I that derives exclusively from T. (Note that the validity of I for C and for P may also be calculated.) Thus, one can have a completely reliable informant with zero validity for T, but a completely valid informant for T must be completely reliable.
3. If two informants are available for each subject in a population, and the data from each of the two are reasonably reliable and have some validity for T, the correlation between the reports of the two informants is the percentage of the total variance that comes from T, plus some portion of the total variance that comes from C and P, depending on how correlated (collinear or overlapping) the two informants' contexts and perspectives are. The correlation might be very low if the informants hold very different perspectives and appraise the subject from two distinct contexts, or the correlation may be quite high if the informants share perspectives and contexts. Thus, one might expect, for example, that the correlation between reports from a mother and father (both parental perspectives from a primarily home context) would be more highly correlated than those from the mother and a teacher (the perspectives of parent versus nonparent and the contexts of home versus school). One might expect the correlation of the reports from two teachers co-teaching a child in the same classroom to be higher than the correlation of the reports from two teachers teaching the same child in successive years (similar perspectives, but different though correlated contexts). The impressions of a clinician evaluating a child in a laboratory setting might have low correlation with those of both parents and teachers (both different contexts and different perspectives), but the reports of two different clinicians, each of whom is following the same protocol, might be expected to be high.

It has been known for almost a century (23, 24) that to increase the reliability of a measure under this model (i.e., reduce the percentage of variance due to E), one obtains M replicate (same T, P, and C) but independent measures (different E) and averages them. This is the rationale that underlies the standard practice of obtaining assays in triplicate and averaging the results, which disattenuates the reliability of a measure. At the same time, two additional actions are needed:

1. To remove extraneous variance due to P, it is necessary to get multiple independent observations from different Ps but from within the same C and average them.
2. To remove extraneous variance due to C, it is necessary to get multiple independent observations from different Cs but from the same P and average them.

The practical challenge is to do both actions simultaneously and to combine the results because in theory that strategy would remove variance due to C and P and increase the proportion of variance due to T (i.e., increase the validity for T). This is not easy to do, however, since, for the most readily available informants, C and P are often confounded. For example, the mother (P) tends to see her child largely in the home (C), whereas the teacher (P) tends to see the child largely in school (C). Moreover, the instruments used by different informants are often insuffi-

FIGURE 2. Example of Appropriate Selection of Multiple Informants Representing Perspectives and Contexts Likely to Influence Expression of Target Characteristics in Assessment of Developmental Psychopathology in Middle Childhood



ciently parallel to each other, and simply averaging scores measured on possibly different scales may be a mistake. To get the most valid measure of T possible from multiple informants, i.e., to remove extraneous variance due to P, C, and E, we need a better approach to put these principles into operation with multiple informants.

A Pragmatic Approach to the Problem of Multiple Informants

Step 1: Specification of the Population and Characteristic of Interest

The first step is to identify the characteristic (T) of interest, the relevant span of time, and the population to which the results are to be applied. What is valid for clinic-referred children, for example, may not be valid for children in the community. A measurement approach valid at age 1 year may no longer be valid at age 5 years. A valid instrument for measuring a transient condition may not be valid for measuring an enduring characteristic over 6 months.

Step 2: Conceptualization of Context and Perspective

Next, the investigators would consider the totality of contexts and perspectives for which there are informants who are likely to provide reliable and to some degree valid data for the target characteristic. Then the challenge is to divide the totality of contexts envisioned into two or more broad categories that are 1) mutually exclusive, 2) broad enough to span contexts, and 3) likely to influence the expression of T as differently as possible, on the basis of one's knowledge of the field. For a young child, these categories might be home versus nonhome. For a teenager, it might be more appropriate to contrast home versus school versus other setting, in the expectation that the expression of the characteristic in settings away from both home and school may be informative. In the same manner, the investigator would divide the totality of perspectives into two or more broad perspectives that are likely to influence informants' reports as distinctively as possible. For a young child, such perspectives might include self, parent, and nonparent; for an older child, they may include self, parent, adult nonparent, and peer. Then the mix-and-match criteria would require contrasting informants with the same perspectives in different contexts and those in the same context with different perspectives.

In general, it can be shown that if there are c contexts and p perspectives (hence cp cells), the minimal number of informants to fill the cells is cp, but the minimal number of informants to satisfy the mix-and-match criterion is c+p−1. Thus, with two contexts and two perspectives (hence four cells), the minimal number of informants is three. With three contexts and two perspectives (hence six cells), the minimal number of informants is four. With three contexts and three perspectives (hence nine cells), the minimal number of informants is five. The difficulties of finding many reliable and valid informants militate against too complex a conceptualization of context and perspective.

This process creates a context-by-perspective grid, within which each possible, reliable, and at least minimally valid informant fits in one or more cells. Informants should be selected to fill at least three cells of the grid in such a way that each pair of contexts is seen in some one perspective, and each pair of perspectives is seen in some one context. By using this mix-and-match strategy, each of the informants is asked for an evaluation of the target characteristic for each child in the sample.

For example, suppose two contexts (home and nonhome) and two perspectives (self and other) were selected for a given research question as shown in the grid in Figure 2.

There are three possible informants in this example: child, teacher, and mother. Note that, by virtue of all cells being filled, the self perspective and other perspective are covered in the nonhome context as well as the home context, and the contrast of nonhome versus home is seen both from the self perspective and from the other perspective. Thus, this choice of informants (even when the two child reports were combined to avoid correlated errors) satisfies the mix-and-match requirement.

Sometimes self versus other, important as that conceptualization of perspective might be, is simply not possible (e.g., with infant subjects). Then we would have to limit perspective to the subspace of other. We might reconceptualize perspective as family versus nonfamily, with "family" including mother, father, siblings, and other relatives, and "nonfamily" including teachers, clinicians, peers, etc. The minimal number of informants necessary would still be three. However, no combination of mother, father, teacher, and clinician would now satisfy the mix-and-match criterion. Yet the most common combinations of

informants in child assessments include these perspectives, as seen in the grid shown in Figure 3.

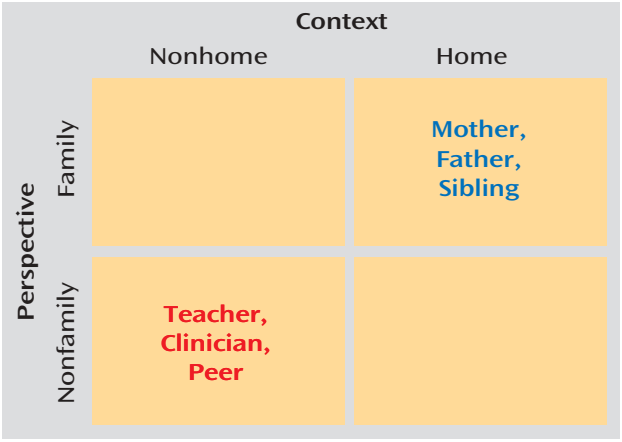
Could we, having mother, father, and teacher readily available, somehow reason backwards and find some conceptualization for which these informants would satisfy the mix-and-match criterion? For example, if the teachers are predominantly women, perhaps we could conceptualize perspective as male versus female. Then mother and father would match on context and contrast on perspective. Mother and teacher would match on perspective and contrast on context. If the conceptualization ill suits the situation (such as the conceptualization of male versus female, which ill suits the current example), the approach will not work (as we will show), but of course serendipity might favor such an attempt.

Clearly there are many possible ways to conceptualize context and perspective, some of which will, and some of which will not, satisfy the mix-and-match criterion. Some conceptualizations will succeed better than others because they reflect a better understanding or knowledge of the field. For example, as shown in Figure 1, one could choose any pair of noncollinear informants to implement triangulation. However, the accuracy of that determination reflects the precision of each informant and the degree of orthogonality between the informants. Thus, a poor choice of informants will result in failed triangulation.

Step 3: Implementation

Once the multiple informants have been selected, subjects are selected from the population of interest, and each informant is asked to report independently on the characteristic of interest for each selected subject. This set of observations is then subjected to principal-component analysis. Principal-component analysis is a mathematical model based on the assumption that the multiple informants' reports are each linear combinations of orthogonal latent variables, as many such latent variables as there are informants per subject. These combinations of variables are labeled in descending order of the percentage of variance explained as the first, second, etc., principal components. Since every informant is selected to provide reliable and at least minimally valid data for T, and since this criterion is the only construct putatively common to all the informants, if the selection of multiple informants is well done, the first principal component should weight each informant's data in the same direction, although not necessarily equally. The factor score on this first factor is a multi-informant estimate of T (which we denote as T*), which is largely free of the effects of C and P and is less affected by E than is any single informant's measure. The second and third factors should correspond to contrasts in context (C*) and perspective (P*) that were built into the choice of informants. Since T*, C*, and P* are linear combinations of data from reliable informants, they should each be reliable. However, T* should be valid for T. Since C* and P* are designed to measure the influence of context and perspective on the expression or reporting of the trait that is independent of the trait itself, C* and P* are constructed to be invalid for T.

FIGURE 3. Example of Inappropriate Selection of Multiple Informants Representing Perspectives and Contexts Likely to Influence Expression of Target Characteristics in Assessment of Developmental Psychopathology in Middle Childhood



It should be emphasized that C* and P* may be interesting variables in their own right. For example, if perspective is structured as self versus other, P* would indicate how discrepant the child's view of him/herself is from the view of others observing the child (e.g., mother and teacher). Such a measure itself may be an indication of a lack of self-awareness in the child or a lack of sensitivity in the adults. That we are here trying to remove the influence of C and P in order to focus on T reflects the goal of this effort to obtain a gold standard measurement of T and does not gainsay the importance of C and P. In fact, an understanding of how both context and perspective influence informants' responses is necessary for successful application of these principles.

Step 4: Validation

The final step of any process of developing a measure is, of course, validation, i.e., demonstration in independent samples from similar populations of the reliability and validity of the proposed multi-informant measure. In the following section, we present some evidence of validation when the process is correctly applied, as well as evidence from the validation attempt that would indicate the process has failed.

An Illustration of the Proposed Approach

Overview

To test the viability and potential utility of the proposed approach to integrating multi-informant data, we used data from three different research projects: the Wisconsin Study of Family and Work (25), the MacArthur Three-City Outcome Study (26), and the Schoolchildren and Their

TABLE 1. Characteristics of Three Studies Whose Data on Developmental Psychopathology in Middle Childhood Were Used to Test the Utility of a Proposed Approach to Integrating Data From Multiple Informants

Characteristic	MacArthur Three-City Outcome Study	Wisconsin Study of Family and Work	Schoolchildren and Their Families Project
Author	Ablow et al. (26)	Essex et al. (25)	Cowan and Cowan (27)
Design	Cross-sectional, case-control design at three study sites	Prospective, longitudinal design involving unselected families from the community	Prospective, longitudinal intervention study involving unselected families from the community
Sample size	120 (67 community subjects and 53 clinical subjects)	275 first-grade children; 264 third-grade children	120 (60 intervention families and 60 comparison families)
Cultural/ethnic composition of sample	87% European-American, 13% minorities	90% European-American, 10% minorities	82% European-American, 18% minorities
Average annual family income	$68,000	$57,000	$72,000
Age (years) and grade of children at assessment(s)	Kindergarten or first grade: mean=5.9, SD=1.0	First grade: mean=6.3, SD=0.7; third grade: mean=8.7, SD=0.8	Prekindergarten: mean=4.8, SD=0.7; kindergarten: mean=5.6, SD=0.7; first grade: mean=6.9, SD=0.8
Child self-report measure	Berkeley Puppet Interview (26)	Berkeley Puppet Interview (26)	Berkeley Puppet Interview (28)
Parent rating scale	MacArthur Health and Behavior Questionnaire (25, 29)	MacArthur Health and Behavior Questionnaire (25, 29)	Child Adaptive Behavior Inventory (27)
Teacher rating scale	MacArthur Health and Behavior Questionnaire (25, 29)	MacArthur Health and Behavior Questionnaire (25, 29)	Child Adaptive Behavior Inventory (27)
Clinical rating scale	—	—	Child-Style Rating System (27)

Families Project (27). The three studies shared major features, specifically a focus on developmental psychopathology in middle childhood (ages 4–8 years) and the use of multiple methods (questionnaires, observations, experimentation) and multiple informants to garner reports on children's symptoms, impairment, and health status. The studies differed, however, in overall design and in the characteristics of the samples. Specifically, the Schoolchildren and Their Families Project and the Wisconsin Study of Family and Work were prospective, longitudinal investigations of children's development that utilized unselected community samples (25, 27). The MacArthur Three-City Outcome Study, by contrast, was a cross-sectional, case-control study in which half the sample was recruited from the community and half from mental health clinics or hospitals. This study was completed to test the reliability and discriminant validity of measures used in the Wisconsin Study of Family and Work. In addition to the references describing the studies, Table 1 provides details about the studies, particularly the range of children's ages at the assessment(s) and the measures used to collect information from multiple informants.

Results

Using the data for first-grade schoolchildren from the Wisconsin Study of Family and Work, we conducted multiinformant principal-component analyses with the mothers', teachers', and children's reports (I) about three different childhood characteristics (T): 1) internalizing symptoms (depression, general anxiety, and separation anxiety), 2) externalizing symptoms (conduct problems, oppositionality, and overt aggression), and 3) academic functioning (academic competence and school engagement). As we discussed earlier, this combination of informants was one that might meet the theory's mix-and-match criterion. The results of these analyses are presented in Table 2.

For each childhood characteristic, the data from all three informants loaded substantially and positively on the first factor, T*, which we interpret as a measure of T. This interpretation is supported by the facts that 1) the second factor (P*) contrasted the mothers' and the teachers' responses with those of the children, and 2) the third factor (C*) contrasted the mothers' and the teachers' responses, with the children's responses (which incorporate both contexts) intermediate between the two and having a near-zero weight. These results capture the mixes and matches that motivated this choice of informants.

With data from the Wisconsin Study of Family and Work and the Schoolchildren and Their Families Project, Table 3 provides the results from a situation in which it is not clear that the three informants have been selected appropriately. Here, mothers', fathers', and teachers' observations of children's internalizing symptoms are used, but the children's self-reports are not.

In the analysis of data from the Wisconsin Study of Family and Work shown in Table 3, the first factor weights each informant in the same direction and thus might conceivably be T*. However, the second factor, which appears to contrast teachers' versus mothers' and fathers' responses, might be either C (home versus nonhome) or P (family versus nonfamily). The third factor contrasts fathers' and teachers' responses versus mothers' responses, which seems inexplicable in terms of either context or perspective. In this case, there is a lack of empirical indication that C and P have been removed and thus that T is a more valid measure. This result is consistent with the theory.

In the analysis of data from the Schoolchildren and Their Families Project shown in Table 3, mothers', teachers', and laboratory-based observations of children's internalizing behavior problems (T) are used to demonstrate further the effects of questionably selected informants. In this example, the third informant is a clinically trained research assistant who provided global ratings of children's

TABLE 2. Principal-Component Analysis Illustrating Appropriate Selection of Multiple Informants in Assessments of Three Characteristics of First-Grade Children in the Wisconsin Study of Family and Work

Characteristic, Informant Group, and Variance Attributable to Factor	Trait[a]	Perspective[b]	Context[c]	Sources of Variability in Informant's Report
	Factor Weight	Factor Weight	Factor Weight	
Internalizing symptoms				
Mothers	0.76	–0.12	–0.64	Other (perspective), home (context)
Teachers	0.69	–0.54	0.49	Other (perspective), school (context)
Children	0.61	0.76	0.25	Self (perspective), home and school (context)
	%	%	%	
Variance attributable to factor	47.1	29.1	23.7	
	Factor Weight	Factor Weight	Factor Weight	
Externalizing symptoms				
Mothers	0.82	–0.18	–0.54	Other (perspective), home (context)
Teachers	0.78	–0.42	0.48	Other (perspective), school (context)
Children	0.60	0.79	0.13	Self (perspective), home and school (context)
	%	%	%	
Variance attributable to factor	54.3	27.9	12.8	
	Factor Weight	Factor Weight	Factor Weight	
Academic functioning				
Mothers	0.88	–0.27	–0.39	Other (perspective), home (context)
Teachers	0.88	–0.29	0.38	Other (perspective), school (context)
Children	0.64	0.77	0.01	Self (perspective), home and school (context)
	%	%	%	
Variance attributable to factor	65.4	24.8	9.7	

[a] Characteristics assessed in the study (internalizing symptoms, externalizing symptoms, academic functioning).
[b] Possible perspectives: self (child) and other (mother, teacher).
[c] Possible contexts: home and school.

TABLE 3. Principal-Component Analysis Illustrating Inappropriate Selection of Multiple Informants in Assessment of Internalizing Symptoms in First-Grade Children in the Wisconsin Study of Family and Work and the Schoolchildren and Their Families Project

Study, Informant Group, and Variance Attributable to Factor	Trait[a]	Perspective[b]	Context[c]	Sources of Variability in Informant's Report
	Factor Weight	Factor Weight	Factor Weight	
Wisconsin Study of Family and Work				
Mothers	0.81	–0.14	–0.58	Mother (perspective), home (context)
Fathers	0.74	–0.48	0.48	Father (perspective), home (context)
Teachers	0.60	0.78	0.18	Teacher (perspective), school (context)
	%	%	%	
Variance attributable to factor	51.7	28.4	19.8	
	Factor Weight	Factor Weight	Factor Weight	
Schoolchildren and Their Families Project				
Mothers	0.67	–0.41	0.79	Parent (perspective), home (context)
Teachers	0.58	0.60	–0.22	Nonparent (perspective), school (context)
Clinicians	–0.53	0.51	0.49	Nonparent (perspective), laboratory/clinic (context)
	%	%	%	
Variance attributable to factor	42.2	31.0	26.8	

[a] Characteristic assessed in the study (internalizing symptoms).
[b] Perspectives for the Wisconsin Study of Family and Work: mother, father, and teacher; perspectives for the Schoolchildren and Their Families Project: parent and nonparent.
[c] Contexts for the Wisconsin Study of Family and Work: home and school; contexts for the Schoolchildren and Their Families Project: home, school, and laboratory/clinic.

TABLE 4. Test-Retest Reliability Coefficients and Effect Sizes for Factor Weights and Mean Raw Scores From Multiple Informants' Reports on Three Characteristics in Community and Clinical Samples of Children in the MacArthur Three-City Outcome Study

Characteristic and Measure	Test-Retest Reliability Coefficient (r)		
	Community Sample (N=29)	Clinical Sample (N=21)	Effect Size (d)
Internalizing symptoms			
Factor weights			
Trait	0.64	0.82	1.3
Perspective	0.62	0.67	0.2
Context	0.78	0.70	0.3
Mean raw scores			
Mothers	0.81	0.85	0.9
Teachers	0.85	0.90	0.8
Children	0.58	0.73	0.7
Externalizing symptoms			
Factor weights			
Trait	0.81	0.96	1.2
Perspective	0.61	0.81	0.1
Context	0.79	0.94	0.1
Mean raw scores			
Mothers	0.82	0.85	1.4
Teachers	0.85	0.97	1.2
Children	0.67	0.70	0.8
Attention deficit hyperactivity disorder symptoms			
Factor weights			
Trait	0.59	0.88	1.2
Perspective	0.50	0.72	0.3
Context	0.72	0.87	0.1
Mean raw scores			
Mothers	0.86	0.90	0.8
Teachers	0.92	0.96	1.4
Children	0.51	0.76	0.4

behaviors after a 2-hour laboratory visit and assessment. Here, none of the factors in the principal-component analysis is immediately identifiable as a measure of T*, C*, or P*. The first factor does not weight the data from all informants in the same direction, and neither the second nor the third factors indicate any relationship to context and perspective. Again, this failure is consistent with the theory.

Evidence of the Reliability and Validity of T, C*, and P**

The examples in Table 2 suggest (but do not prove) that the panel of informants consisting of mothers, teachers, and children will enable us to compute a measure of T that 1) combines each informant's observation of that trait and 2) is relatively free from variance attributable to C and P. On the basis of the theory, it is predicted that T*, C*, and P* should all be reliable measures and that T* should be valid for T while C* and P* should not be valid for T. How well do these measures fulfill those predictions? For this purpose, data from the MacArthur Three-City Outcome Study were used, and T*, C*, and P* were calculated as in the Wisconsin Study of Family and Work described earlier.

Table 4 presents reliability coefficients for T*, C*, and P* separately in the clinical and community samples in the MacArthur Three-City Outcome Study. Despite the limitation in the range of values for the clinical and community samples (which usually depresses reliability), T*, C*, and P* all had satisfactory test-retest reliability, as did the separate scores of the three types of informants.

Table 4 also presents the effect size (standardized mean difference between the means for the clinical and community samples) indicating the discriminant validity for the presence/absence of clinical problems. It can be seen that T* is valid (an effect size of 0.8 is typically considered large) (30) but that P* and C*, as predicted, are not (an effect size of 0.2 is typically considered small). For internalizing symptoms, the effect size of T* is larger than the effect sizes for any of the categories of informant. For externalizing symptoms and ADHD, the effect size for T* is not the largest, but it must be remembered that referral of children, and thus their inclusion in the clinical sample, is typically based on the mother's or teacher's reports.

Developmental Issues and Multi-Informant Estimates of T, C, and P

In a final set of analyses, data from the Schoolchildren and Their Families Project and the Wisconsin Study of Family and Work were used to give additional insights into the validity of T*. We examined three issues. First, are the results replicated from one study or site to another in community samples (i.e., are the results generalizable across communities)? Second, does the proportion of variance attributable to T change as a function of children's age? The age ranges in these studies were very narrow, but in general one would expect that as the child matures, obtaining a gold standard measure would become easier and the percent of variance attributable to T would thus increase. Finally, how does each informant's relative contribution (loadings) to the multi-informant measures of T, C, and P change as a function of children's age? As the child matures, gaining in introspection and communication skills, one would expect that the weight placed on the child's report would increase and the weight placed on the mother's and teacher's reports would decrease.

Table 5 presents the percent of variance accounted for by each multi-informant factor at four different grade levels during middle childhood: prekindergarten, kindergarten, and first grade in the Schoolchildren and Their Families Project, and first and third grades in the Wisconsin Study of Family and Work. For generalizability, the crucial comparison is between the first-grade samples at both sites. In both studies, the expected increase is seen in the percentage of total variance accounted for by T. Table 6 presents a closer look at changes in the mothers', teachers', and children's weights in T* at each of these same time points. Again, the crucial comparison between the two first-grade samples indicates remarkable agreement between the two studies. Moreover, in most cases (reports on externalizing symptoms in the Schoolchildren and Their Families Project is the exception), the weight on the children's reports tends to increase with the age of the child,

TABLE 5. Percent of Variance Attributable to Trait, Perspective, and Context Factors in Multiple Informants' Reports on Three Characteristics in Subjects in the Schoolchildren and Their Families Project and the Wisconsin Study of Family and Work

	Variance Attributable to Factor (%)				
	Subjects in the Schoolchildren and Their Families Project			Subjects in the Wisconsin Study of Family and Work	
Characteristic and Factor	Prekindergarten Children	Kindergarten Children	First-Grade Children	First-Grade Children	Third-Grade Children
Internalizing symptoms					
Trait	42.7	45.1	47.9	47.2	49.7
Perspective	30.3	28.6	25.7	25.1	21.9
Context	27.0	26.3	26.4	27.7	28.4
Externalizing symptoms					
Trait	52.3	51.9	53.4	54.3	53.5
Perspective	26.5	30.3	23.6	27.9	24.1
Context	21.2	17.8	23.0	17.8	22.4
Attention deficit hyperactivity disorder symptoms					
Trait	55.2	57.0	59.9	60.6	65.8
Perspective	28.2	25.6	24.0	23.6	20.1
Context	16.6	17.4	16.1	15.8	14.1

TABLE 6. Factor Weights for the Trait Factor in Multiple Informants' Reports on Three Characteristics in Subjects in the Schoolchildren and Their Families Project and the Wisconsin Study of Family and Work

	Factor Weight				
	Subjects in the Schoolchildren and Their Families Project			Subjects in the Wisconsin Study of Family and Work	
Characteristic and Informant Group	Prekindergarten Children	Kindergarten Children	First-Grade Children	First-Grade Children	Third-Grade Children
Internalizing symptoms					
Mothers	0.83	0.80	0.74	0.73	0.55
Teachers	0.79	0.78	0.79	0.78	0.78
Children	0.58	0.58	0.63	0.62	0.79
Externalizing symptoms					
Mothers	0.85	0.85	0.80	0.82	0.77
Teachers	0.83	0.81	0.78	0.75	0.74
Children	0.61	0.64	0.62	0.60	0.70
Attention deficit hyperactivity disorder symptoms					
Mothers	0.82	0.86	0.81	0.85	0.80
Teachers	0.81	0.79	0.82	0.82	0.84
Children	0.59	0.66	0.73	0.70	0.78

reflecting the expected developmental growth in valid self-reports.

Summary and Discussion

Despite widespread agreement that the assessment of certain subject characteristics requires multiple informants (e.g., self, parent, teacher), convergence of data from the various informants is almost never achieved. In the absence of gold standard measures, scientists and clinicians must contend with low levels of agreement between informants. Although our knowledge about the sources of the variability in informants' reports is limited, high levels of variability consistently imply that an informant's report reflects the separate and combined influences of 1) the actual characteristic (e.g., traits, symptoms, competencies), 2) the context (or situations) in which the subject is observed, 3) the perspectives (or biases) of the informant, and 4) error of measurement. Although the illustration in the current study focused on child characteristics, the foregoing discussion has proposed and tested a general approach to the integration of multi-informant data that is intended to improve validity in measurement of a variety of characteristics.

The proposed approach is predicated on the following general assumptions and procedures:

1. A reliable and reasonably valid informant's report (I) comprises information on the trait or characteristic in question (T), some contribution from the context in which that informant is likely to observe the subject (C), a contribution from the perspective from which that informant views the subject (P), and random error (E). Each of these pieces of information, or sources of variance, can be defined as orthogonal latent variables.
2. For this three-dimensional model, we need at least three informants, each carefully selected to report reliable information about the specific characteristic, knowing that no one informant has all the pertinent information.
3. In selecting our informants, we would not choose informants likely to give collinear (highly correlated)

reports because they would simply reproduce the same incomplete information.

4. Rather, we would try to select informants likely to give orthogonal (valid, but not redundant) reports, in such a way as to have the flaws (i.e., variability in the data that is not linked to the target characteristic) in one informant's data "corrected" by other informants.
5. Instead of asking, "How many informants do we need, and how do we combine their reports?" we suggest that the question should rather be, "How do we select informants in such a way that the imperfections in one informant's reports are corrected by another's reports?"
6. To reduce the influence of perspective (P) and context (C), one triangulates the data by using a mix-and-match strategy, in which specific selected contexts are viewed from the same perspective and selected perspectives are viewed in the same context. By choosing informants to implement this mix-and-match strategy, one structures the data from multiple informants in such a way that principal-component analysis will yield a gold standard measure as the first principal component, T*, and measures of the contrasts in context and perspectives as the second and third principal components (C* and P*).
7. If the theory and implementation are correct, T* will be a reliable and valid measure of T. C* and P* are both reliable measures, but they are invalid for T. Thus, the model removes from T* the sources of error about T that are represented by C and P in the individual informants' reports.

It is important to note that with this model we both 1) explain why past attempts using principal-component analysis or factor analysis often did not work and 2) show that such analyses do work with well-designed choices of multiple informants.

The approach defined here does not underestimate the difficulty of making informed, careful choices of multiple informants in designing effective research. For each characteristic there are many choices of C and P, and the choice of an appropriate grid and of the appropriate informants must be based on a deep understanding of the characteristic and the subjects in question. More important yet, as the subject develops or ages, the definitions and choices of C and P might well differ. For example, as a child enters the teen years, the self (the child) might become a more important informant than the parent. The parent might become irrelevant as an informant when the child becomes an independent adult. Spouses or important others might then become more important. Similarly, school is a major context for young children, but it is irrelevant as they reach adulthood, when it is likely to be replaced by the workplace. In another example, when assessing characteristics of patients with Alzheimer's disease in the early stages, the researcher would likely use self versus other perspectives and familiar versus unfamiliar contexts. Thus, the Alzheimer's disease patient, the caregiver, and the clinician might be good choices of informants. In later stages, with the cognitive decline of the patient, the percent variance accounted for by the patient's reports will decrease, and the weight on the patient's response will as well, until eventually the Alzheimer's disease patient cannot be included as an informant. At that point, the contexts would likely remain familiar (home) versus unfamiliar (clinic or laboratory), but the perspectives might change to important other (e.g., caregiver) versus stranger (e.g., clinical observer or tester).

The ultimate goal of this new approach to utilizing data from multiple informants is to provide more valid assessment of disorder in studies related to mental health problems in children. By disaggregating the variability attributable to traits, contexts, perspectives, and error, we can augment dramatically our capacity for usefully integrating informants' views and thereby advance the utility of psychiatric epidemiology and clinical trials. It is our hope that such an advance might substantively and helpfully inform emerging knowledge of the origins and prevention of human psychopathology.

Received Jan. 31, 2002; revision received Jan. 15, 2003; accepted Feb. 2, 2003. From the Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, Calif.; the Department of Psychology, University of Oregon, Eugene, Ore.; the Department of Psychiatry, University of Wisconsin, Madison, Wisc.; the School of Public Health and the Institute of Human Development, University of California, Berkeley, Calif.; and the Department of Psychiatry, Western Psychiatric Institute, University of Pittsburgh, Pittsburgh. Address reprint requests to Dr. Kraemer, Department of Psychiatry and Behavioral Sciences, Stanford University, 401 Quarry Rd., MC 5717, Stanford, CA 94305; HCK@Stanford.edu (e-mail).

Supported by the John D. and Catherine T. MacArthur Foundation Research Network on Psychopathology and Development (Dr. Kupfer, Chair) and NIMH grants MH-44340 and MH-52354 (Dr. Essex).

The authors thank Drs. David Offord, Alan Kazdin, Phil Cowan, Carolyn Cowan, and the other members of the John D. and Catherine T. MacArthur Foundation Research Network on Psychopathology and Development for their contributions.

References

1. National Institutes of Health: NIH Policy and Guidelines on the Inclusion of Children as Participants in Research Involving Human Subjects, March 6, 1998 (http://grants2.nih.gov/grants/guide/notice-files/not98-024.html)
2. Pedhazer EJ, Schmelkin L: Measurement, Design, and Analysis: An Integrated Approach. Hillsdale, NJ, Lawrence Erlbaum Associates, 1991
3. Angold A, Costello JE: The relative diagnostic utility of child and parent reports of oppositional defiant behaviors. Int J Methods Psychiatr Res 1996; 6:253–259
4. Piacentini JC, Cohen P, Cohen J: Combining discrepant information from multiple sources: are complex algorithms better than simple ones? J Abnorm Child Psychol 1992; 20:51–63
5. Silverman WK, Eisen AR: Age differences in the reliability of parent and child reports of child anxious symptomatology using a structured interview. J Am Acad Child Adolesc Psychiatry 1992; 31:117–124

6. Achenbach TM, McConaughy SH, Howell CT: Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. Psychol Bull 1987; 101:213–232
7. Hinshaw SP, Han SS, Erhardt D, Huber A: Internalizing and externalizing behavior problems in preschool children: correspondence among parent and teacher ratings and behavior observations. J Clin Child Psychol 1992; 21:143–150
8. Kazdin AE (ed): Informant Variability in the Assessment of Childhood Depression. New York, Plenum, 1994
9. McConaughy SH, Stanger C, Achenbach TM: Three-year course of behavioral/emotional problems in a national sample of 4- to 16-year-olds, I: agreement among informants. J Am Acad Child Adolesc Psychiatry 1992; 31:932–940
10. Carbonneau R, Rutter M, Simonoff E, Silberg JL, Maes HH, Eaves LJ: The Twin Inventory of Relationships and Experiences (TIRE): psychometric properties of a measure of the non-shared and shared environmental experiences of twins and singletons. Int J Methods Psychiatr Res 2001; 10:72–85
11. Offord DR, Boyle MH, Racine Y, Szatmari P, Fleming JE, Sanford M, Lipman EL: Integrating assessment data from multiple informants. J Am Acad Child Adolesc Psychiatry 1996; 35:1078–1085
12. Shoda Y, Mischel W, Wright JC: Intuitive interactionism in person perception: effects of situation-behavior relations on dispositional judgments. J Pers Soc Psychol 1989; 56:41–53
13. Shoda Y, Mischel W, Wright JC: Links between personality judgments and contextualized behavior patterns: situation-behavior profiles of personality prototypes. Social Cognition 1993; 11:399–429
14. Achenbach TM: Taxonomy and comorbidity of conduct problems: evidence from empirically based approaches. Dev Psychopathol 1993; 5:51–64
15. Boyle MH, Offord DR, Racine YA, Szatmari P, Sanford M, Fleming JE: Interviews versus checklists: adequacy for classifying childhood psychiatric disorder based on adolescent reports. Int J Methods Psychiatr Res 1996; 6:309–319
16. Verhulst FC: Recent developments in the assessment and diagnosis of child psychopathology. Eur J Psychol Assess 1995; 11: 203–212
17. Kolko DJ, Kazdin AE: Emotional/behavioral problems in clinic and nonclinic children: correspondence among child, parent and teacher reports. J Child Psychol Psychiatry 1993; 34:991–1006
18. MacLeod RJ, McNamee JE, Boyle MH, Offord DR, Friedrich M: Identification of child psychiatric disorder by informant: comparisons of clinic and community samples. Can J Psychiatry 1999; 44:144–150
19. Infant Health and Development Program: enhancing the outcomes of low birth weight, premature infants: a multisite, randomized trial. JAMA 1990; 263:3035–3042
20. MTA Cooperative Group: A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder: the Multimodal Treatment Study of Children With ADHD. Arch Gen Psychiatry 1999; 56:1073–1086
21. Swanson JM, Kraemer HC, Hinshaw SP, Arnold E, Connors CK, Abikoff HB, Clevenger W, Davies M, Elliott GR, Greenhill LL, Hechtman L, Hoza B, Jensen PS, March JS, Newcorn JH, Owens EB, Pelham WE, Schiller E, Severe JB, Simpson S, Vitello B, Wells K, Wigal T, Wu M: Clinical relevance of the primary finding of the MTA: success rates based on severity of ADHD and ODD symptoms at the end of treatment. J Am Acad Child Adolesc Psychiatry 2001; 40:168–179
22. Conners CK, Epstein JN, March JS, Angold A, Wells KC, Klaric J, Swanson JM, Arnold LE, Abikoff HB, Elliott GR, Greenhill LL, Hechtman L, Hinshaw SP, Hoza B, Jensen PS, Kraemer HC, Newcorn J, Pelham WE, Severe JB, Vitiello B, Wigal T: Multimodal treatment of ADHD in the MTA: an alternative outcome analysis. J Am Acad Child Adolesc Psychiatry 2001; 40:159–167
23. Brown W: Some experimental results in the correlation of mental abilities. Br J Psychol 1910; 3:296–322
24. Spearman C: Correlation calculated from faulty data. Br J Psychol 1910; 3:271–295
25. Essex MJ, Klein MH, Miech R, Smider NA: Timing of initial exposure to maternal major depression and children's mental health symptoms in kindergarten. Br J Psychiatry 2001; 170: 151–156
26. Ablow JC, Measelle JR, Kraemer HC, Harrington R, Luby J, Smider N, Dierker L, Clark V, Dubika B, Heffelfinger A, Essex MJ, Kupfer DJ: The MacArthur Three-City Outcome Study: evaluating multi-informant measures of young children's symptomatology. J Am Acad Child Adolesc Psychiatry 1999; 38:1580–1590
27. Cowan CP, Cowan PA: Working with couples during stressful transitions, in The Family on the Threshold of the 21st Century: Trends and Implications. Edited by Dreman S. Mahwah, NJ, Lawrence Erlbaum Associates, 2000, pp 17–47
28. Measelle JR, Ablow JC, Cowan PA, Cowan CP: Assessing young children's views of their academic, social, and emotional lives: an evaluation of the self-perception scales of the Berkeley Puppet Interview. Child Dev 1998; 69:1556–1576
29. Boyce WT, Quas J, Alkon A, Smider NA, Essex MJ, Kupfer DJ: Autonomic reactivity and psychopathology in middle childhood. Brit J Psychiatry 2001; 179:144–150
30. Cohen J: Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ, Lawrence Erlbaum Associates, 1988