

Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications, and future directions

Thomas M. Achenbach,¹ Andreas Becker,² Manfred Döpfner,³ Einar Heiervang,⁴ Veit Roessner,² Hans-Christoph Steinhausen,⁵ and Aribert Rothenberger²

¹Department of Psychiatry, University of Vermont, Burlington, USA; ²Child and Adolescent Psychiatry, University of Goettingen, Germany; ³Child and Adolescent Psychiatry, University of Cologne, Germany; ⁴Child and Adolescent Psychiatry, University of Bergen, Norway; ⁵Child and Adolescent Psychiatry, University of Zürich, Switzerland

Around the world, cultural blending and conflict pose challenges for assessment and understanding of psychopathology. Economical, evidence-based, culturally robust assessment is needed for research, for answering public health questions, and for evaluating immigrant, refugee, and minority children. This article applies multicultural perspectives to behavioral, emotional, and social problems assessed on dimensions describing children's functioning, as rated by parents, teachers, children, and others. The development of Achenbach System of Empirically Based Assessment (ASEBA) and Strengths and Difficulties Questionnaire (SDQ) forms and their applications to multicultural research are presented. A primary aim of both questionnaires is to identify children at high risk of psychiatric disorders and who therefore warrant further assessment. The forms are self-administered or administered by lay interviewers. ASEBA problem items are scored on 6 DSM-oriented scales and 3 broader band scales, plus 8 syndromes derived statistically as taxonomic constructs and supported by uniform confirmatory factor analyses of samples from many populations. Comparisons of ASEBA scale scores, psychometrics, and correlates are available for diverse populations. SDQ forms are scored on one broad-band scale and 5 a priori behavioral dimensions supported by data from various populations. For both instruments, factor analyses, psychometrics, and correlates are available for diverse populations. The willingness and ability of hundreds of thousands of respondents from diverse groups to complete ASEBA and SDQ forms support this approach to multicultural assessment. Although particular items and scales may have differential relevance among groups and additional assessment procedures are needed, comparable results are found in many populations. Scale scores vary more within than between populations, and distributions of scores overlap greatly among different populations. Ratings of children's problems thus indicate more heterogeneity within populations than distinctiveness between populations. Norms from multiple populations can be used to compare children's scores with relevant peer groups. Multicultural dimensional research can advance knowledge by diversifying normative data; by comparing immigrant children with nonimmigrant compatriots and with host country children; by identifying outlier findings for elucidation by emic research; and by fostering efforts to dimensionalize DSM-V diagnostic criteria. **Keywords:** Multicultural, psychopathology, assessment, dimensional, informant ratings, cross-cultural, Child Behavior Checklist, rating scales, trans-cultural, Strengths and Difficulties Questionnaire, dimensional assessment. **Abbreviations:** ASEBA: Achenbach System of Empirically Based Assessment; SDQ: Strengths and Difficulties Questionnaire; CBCL: Child Behavior Checklist; TRF: Teacher's Report Form; YSR: Youth Self-Report.

The 21st century presents abundant opportunities and challenges for those who are committed to helping the world's children. (For brevity, we use 'children' to include adolescents.) Research and services related to children's problems are advancing. Computers and the internet facilitate access to knowledge and technology throughout the world. The blending of different cultures creates vibrant new combinations and perspectives. On the other

hand, millions of immigrants and refugees challenge the mental health, medical, social welfare, educational, and forensic systems of host societies. Societies populated by multiple native-born cultural groups also face challenges in equitably providing for the different groups. Such challenges are greatly intensified by both the fear and reality of clashes between various cultural groups.

The mixing of cultural groups and the occasional clashes between them make it imperative for us to expand our perspectives beyond our own personal backgrounds. This is essential both for working with individual children of different backgrounds and for addressing the public health needs of

Conflict of interest statement: T.M. Achenbach is author or co-author, and President of the nonprofit Research Center for Children, Youth, and Families, Inc., which publishes the ASEBA and from which he receives remuneration. M. Döpfner receives a research grant for CBCL, TRF and YSR.

various societies. In particular, growing recognition of child mental health needs is increasing the demand for cost-effective evidence-based assessment of children in both developed and developing societies.

As charged by the editors of the *Journal of Child Psychology and Psychiatry*, we focus mainly on the two sets of instruments that have been used most extensively to assess child psychopathology in diverse cultural contexts. These instruments can be readily used by researchers and practitioners in many societies. Findings with these instruments have already provided a great deal of information about similarities and differences between problems reported for children from many different backgrounds. After we present the conceptual frameworks for these instruments, details of the instruments themselves, and findings obtained with them, we will integrate the findings to form a foundation on which to base future advances in research and services.

Multicultural perspectives and dichotomies

Cross-cultural research has a long history. However, it has tended to focus on differences between cultures as if each culture is 'internally homogeneous and externally distinctive' (Hermans & Kempen, 1998, p. 1119). This view of cultures has engendered a variety of dichotomies for simplifying the welter of differences that can be identified between various groups. Such dichotomies include the 'West versus the Rest' (Hermans & Kempen, 1998) and 'individualism versus collectivism' (Triandis, 1989). At a more primitive level, most humans tend to simplify their myriad differences by imposing a dichotomy that can be called 'we versus they.' Virtually any discernible differences can become criteria for discriminating the we from the they, usually to the detriment of the 'they.'

To advance understanding of children's problems, we must remind ourselves that, as members of the same species, all humans follow similar sequences of biological, cognitive, and linguistic development. Yet, within this common developmental framework, there are marked individual differences in developmental rates, temperament, and adaptive success. Cultural factors help to shape the environments in which children develop. However, complex interplays of these cultural factors with other aspects of children's environments and with biological factors may contribute as much to individual differences within cultural groups as to differences between groups.

Multicultural perspectives and dimensions

To take account of individual differences, group differences, and the interplay of individual and group differences, we use the term *multicultural* in preference to 'cross-cultural.' As we use it, *multicultural*

encompasses both individual and group differences that can be viewed in terms of variations along quantitative dimensions. Multicultural perspectives can help researchers and practitioners avoid imposing artificial dichotomies on people. When viewed from multicultural perspectives, identification of both similarities and differences requires measurement of the characteristics of many individuals from many populations. After measurements have been obtained for normative samples of individuals from various populations, similar measurements of new individuals enable us to compare them with the norms for relevant populations.

The following concrete example illustrates this point. Suppose that the mean height of 8-year-old children in Society A exceeds the mean height of 8-year-olds in Society B. Furthermore, the mean height in Society B exceeds that in Society C. Although the mean heights differ, there would be a great deal of overlap between the heights of 8-year-olds in Societies A, B, and C. Consequently, it would be wrong to think that every 8-year-old in Society A is taller than every 8-year-old in Societies B and C, or that tall stature is an intrinsic characteristic of children in Society A, while short stature is an intrinsic characteristic of children in Society C. On the contrary, the differences in height *within* Societies A, B, and C are likely to be considerably greater than the differences *between* the mean heights in Societies A, B, and C.

By measuring the heights of representative samples of children in Societies A, B, and C, we can establish age- and gender-specific norms for height in each society. Age- and gender-specific norms would enable us to identify children who are small enough to be candidates for interventions such as growth hormones and nutritional supplements. If the distributions of heights in Societies A, B, and C are found to be quite similar, these distributions can be pooled to form a single set of norms for use in all three societies. Height can be uniformly measured with a single standardized instrument in all societies.

The measurement of child psychopathology is much more complex, however, because the phenomena to be measured cannot be defined in terms of physical units. Instead, measurement of child psychopathology involves judgments by multiple people that certain behavioral, emotional, cognitive, and/or social characteristics are sufficiently detrimental to warrant professional help. At this stage of our knowledge, it is not realistic to expect universal agreement on a single definition of child psychopathology. In lieu of a single definition, we focus on behavioral, emotional, and social problems that (a) can be reported by parents, teachers, and children, and (b) are found to be associated with referral for mental health services and for similar kinds of help in multiple societies. Other kinds of problems, including those measured with cognitive and

biomedical instruments, may be equally important but are beyond the scope of this article.

Although far simpler than measurement of children's behavioral, emotional, and social problems, even the measurement of height is affected by variations in the phenomena being measured and in the measuring instruments themselves. For example, the results may differ when children are measured in prone versus standing positions, when measurements are made at different times of day or on different days, and when different measuring instruments are used. Measurement 'error,' i.e., variance, is a basic fact of life in all measurement procedures. Measurement variance arises not only from mistakes in performing and recording measurements, but from differences in measurement procedures, measurement occasions, and other factors, many of which may not be precisely identified. Despite the inevitability of measurement variance, quantification can greatly increase the precision with which we assess and understand the target phenomena. By quantifying psychopathology in terms of dimensional aggregations of scores for co-occurring problems, we can measure variance in reported problems much more precisely than if each specific problem and aggregation of problems must be categorized as present versus absent.

Each dimension can include many scores. Consequently, differences found between individual children and between groups can be conceptualized in relation to a range of possible scores on each dimension. Rather than viewing all children of one cultural group as categorically different from all children of another group, we can view each child in terms of quantified profile patterns consisting of scores on dimensions that can be compared with distributions of scores in normative samples drawn from particular cultural groups. The use of quantified dimensions of problems to evaluate children in relation to norms for particular cultural groups is analogous to the evaluation of children's heights in relation to norms for their age and gender.

Etic and emic viewpoints

Our multicultural approach to psychopathology is consistent with the *etic* approach to studies of cultural variations articulated by the linguist Kenneth Pike (1954, 1967). Pike coined the term *etic* by shortening the linguistic term *phonetic*, which refers to standard systems for representing the total repertoire of sounds used in all human languages. Like linguists' phonetic systems, the multicultural approach to psychopathology uses similar standardized assessment methods in different cultures.

As a contrast to the *etic* approach, Pike coined the term *emic* by shortening the linguistic term *phonemic*, which refers to the sounds that are meaningful in a particular language. The *etic* approach is in no way incompatible with the *emic*

approach to understanding the meaning that particular characteristics have in a particular culture. On the contrary, the *etic* use of standardized multicultural assessment can reveal differences between cultural groups that can then be investigated with *emic* methods to identify differences in the etiologies and meanings of particular characteristics found in particular cultural groups. As an example, suppose that *etic* methods for assessing certain problems yield significantly higher scores for boys than girls in most but not all cultural groups. *Emic* methods can then be used to investigate reasons for the exceptions to the tendency for boys to obtain higher scores than girls in most cultural groups.

Categorically based and dimensionally based assessment of psychopathology

Psychopathology can be defined as the systematic study of abnormal behavior, experiences, and cognitions (Sims et al., 2000). Descriptive psychopathology describes and aggregates abnormal behaviors, experiences, and cognitions that are observed or are reported by patients or by proxies. Both observable behaviors and internal experiences and cognitions may be regarded as symptoms if they clearly reflect abnormal functioning.

Research on child psychopathology in various cultural groups has mainly employed diagnostically based and empirically based assessment methods (Bird, 1996). Diagnostically based methods have tended to assess child psychopathology in terms of categorical, yes-versus-no decisions about whether diagnostic criteria are met for particular disorders. Empirically based methods, on the other hand, have tended to assess psychopathology in dimensional and other quantitative terms. Although there are important points of contact between the different approaches, we preface our review by outlining contrasts between the categorical versus dimensional aspects.

Categorically based assessment

Categorically based assessment stems from nosological models that specify particular problems as being symptoms of disorders. The disorders are classified according to diagnostic categories. The prevailing nosologies for behavioral, cognitive, and emotional disorders are the World Health Organization's (1992) *International Classification of Diseases* (the ICD) and the American Psychiatric Association's (1994) *Diagnostic and Statistical Manual* (the DSM). Categorically based assessment can be viewed as proceeding mainly from 'the top down,' because it starts 'at the top' with the constructs embodied in a nosology. After agreeing on the nosological constructs, the authors of the nosology formulate criteria for determining who has each disorder that is

represented by the nosological constructs. Field trials have been used to test and revise diagnostic criteria for some nosological constructs.

The most common research method for obtaining categorically based data on child psychopathology consists of standardized diagnostic interviews for children and their parents. The interviews are designed to operationalize the criteria for the DSM and ICD nosological constructs mostly by asking questions whose answers are coded as indicating whether or not criteria for diagnoses are met. Various versions of a highly structured standardized interview, the Diagnostic Interview Schedule for Children (DISC; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000), have been administered to epidemiological samples in several cultures (reviewed in detail by Achenbach & Rescorla, 2007b). However, to our knowledge, no published studies report direct statistical comparisons of findings from similar versions of the DISC administered to parallel epidemiological samples from two or more cultures.

Less structured diagnostic interviews have also been developed to operationalize diagnostic criteria. Examples include the Parental Accounts of Symptoms (PACS; Taylor et al., 1986), the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS; Ambrosini, 2000), and the Parent Interview for Child Symptoms (PICS; Ickowicz et al., 2006). We know of no published studies reporting direct statistical comparisons between findings from different cultures for these interviews, either. However, statistical comparisons of diagnoses made with the Development and Well-Being Assessment (DAWBA; Goodman, Ford, Richards, Gatward, & Meltzer, 2000) have been published for samples from the United Kingdom and the municipality of Taubaté, Brazil (Fleitlich-Bilyk & Goodman, 2004).

Dimensionally based assessment

In contrast to categorically based assessment, dimensionally based assessment is characterized more by a 'bottom-up' approach. In this approach, assessment proceeds from 'the bottom up' by obtaining scores for specific descriptors of children's functioning. The scores for specific descriptors are then aggregated into scales for measuring psychopathology and other aspects of functioning. Each child's scale scores can be compared with scores for normative samples in order to evaluate the degree of deviance indicated by the child's scores.

Clinical questionnaires are often used to obtain dimensional data. Such questionnaires have various advantages and disadvantages (Conners, 1998). Advantages include the ability of questionnaires to obtain ratings that draw on experience with a child over extended time intervals and diverse situations. Even rare and infrequent behaviors can be assessed that may be missed by interviews. Questionnaires

are cheap and require little or no professional time to administer. If normative data are available, each child can be evaluated in terms of deviance from what is reported for representative samples of peers. This is helpful for judging whether a child's problems are within or beyond the normative range at initial assessment and again following interventions. Ratings of children's behavior by parents, teachers, and significant others have substantial ecological importance regardless of accuracy or reliability. Finally, ratings permit the quantification of qualitative aspects of behavior in children that is not readily assessed by other means.

Questionnaires also have disadvantages. They may be subject to systematic errors, such as leniency or severity in ratings, halo effects resulting from positive or negative slants toward certain sets of items, logical errors, contrast errors owing to comparisons with particular other children, and recency effects stemming from recent episodes of behavior. Questionnaires are also limited to obtaining the respondents' perspectives on the particular questions that are posed. Furthermore, the respondents' subjective experiences may not be explored, direct observations are not obtained, and misunderstandings may not be clarified. Finally, questionnaires are not sufficient for making clinical diagnoses. Thus, good clinical practice requires a combination of assessment instruments including questionnaires, observations, tests, and interviews, although all methods are vulnerable to multiple sources of error.

As charged by the editors, we review the two sets of dimensional assessment instruments for which the most multicultural findings have been published. These are the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Edelbrock, 1983, 1986, 1987; Achenbach, 1991a, b, c; Achenbach & Rescorla 2001, 2007a) and the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997, 2001; Goodman, Meltzer, & Bailey, 1998; Rothenberger & Woerner, 2004). We limit our review to the parent, teacher, and self-report versions of these instruments that have generated the most research in different cultures, although versions also exist for other ages and other informants.

The ASEBA and SDQ instruments are similarly designed to obtain ratings of children's problems and positive characteristics that are then summed to yield scale scores. Both questionnaires can be used as general screens to identify high-risk children who warrant more careful assessment for disorders. ASEBA and SDQ forms are brief and can be self-administered or administered by lay interviewers in many contexts. Their brevity, ease of administration, and translations into many languages have facilitated their use with diverse cultural groups. Table 1 lists cultural groups for which ASEBA and/or SDQ findings have been published as of this writing.

Table 1 Cultural groups for which ASEBA and/or SDQ findings have been published

^a African American	^a Ghanaian	^a Palestinian
^a American Indian (Native American)	^a Greek	^a Peruvian
^a Armenian	^a Guatemalan	^a Philippino
Australian	^a Hong Kong Chinese	Portuguese
^a Austrian	Icelandic	^a Puerto Rican
^b Bangladeshi	^a Indian	^a Quebeccois
^a Belgian	^a Indonesian	^a Romanian
Brazilian	Iranian	Russian
British	Iraqi	^a Saudi Arabian
^a Bulgarian	^a Irish	^a Serbian
^a Cambodian	^a Israeli	^a Seychelles
^a Canadian	Italian	^b Somalian
^a Caucasian American	^a Jamaican	Sri Lankan
^a Chilean	^a Japanese	Spanish
^a Chinese	^a Kenyan	Swedish
^b Congolese	^a Korean	^a Swiss
Croatian	^a Kurdish	^a Taiwanese
^a Cuban	^a Latino	^a Thai
^a Czech	^a Lebanese	Turkish
Danish	^a Lithuanian	^a Ukrainian
Dutch	^a Malaysian	^a Venezuelan
^a Ethiopian	^a Mauritian	Vietnamese
Finnish	^a Mexican	^b Zambian
Flemish	^a Moroccan	
French	New Zealander	
German	Norwegian	
	^a Pacific Islanders	

^aASEBA only.^bSDQ only.

Differences from diagnostic interviews. Both the ASEBA and SDQ differ from diagnostic interviews in the following ways:

1. The basic ASEBA and SDQ assessment data consist of 0-1-2 ratings of descriptive items by people who see children in particular contexts and by the children themselves. By contrast, psychiatric interview data typically consist of parents' and/or children's yes/no answers to trained interviewers' questions about diagnostic criteria.
2. The ASEBA and SDQ ratings are summed to yield scale scores. By contrast, nosological decision rules are applied to the yes/no answers to psychiatric interviewers' questions in order to determine whether criteria for diagnoses are met.
3. Because respondents rate all items on ASEBA and SDQ forms, a standardized database is available for each child rated on a particular form. By contrast, 'skip-outs' allow questions to be selectively omitted from psychiatric interviews, thereby yielding different databases for different children.
4. The ASEBA and SDQ are not designed to operationalize nosological criteria. However, cutpoints on distributions of scale scores and top-down selections of items can be used to provide cross-walks with diagnoses. By contrast, psychiatric interviews are designed to operationalize nosological criteria by using extensive questions to determine whether specific diagnostic criteria are met.

Differences between the ASEBA and SDQ. Despite similarities in the ASEBA and SDQ and in how they both differ from diagnostic interviews, the ASEBA and SDQ differ from each other in the following ways:

1. Since its inception (Achenbach, 1966), the ASEBA has been designed to identify syndromes of co-occurring problems that can serve as taxonomic constructs for psychopathology. ASEBA forms are therefore scored on syndrome scales that embody statistically identified sets of co-occurring problems, although ASEBA forms are also scored on other kinds of scales, including top-down DSM-oriented scales. The SDQ, by contrast, has been designed to tap dimensions designated as Conduct Problems, Emotional Symptoms, Hyperactivity, Peer Relationships, and Prosocial Behavior (Goodman, 1997, p. 581). After these dimensions were suggested by factor analyses of a modified version of Sir Michael Rutter's (1967) rating scale, Goodman wrote five SDQ items to measure each dimension. The SDQ scales themselves were thus constructed in a top-down way, albeit on the basis of bottom-up findings with the Rutter scale.
2. Ranging from 105 problem items on the YSR to 120 on the CBCL and TRF, ASEBA problem items are more differentiated than the 20 SDQ problem items. All ASEBA problem items and most SDQ problem items are rated so that 0 = absence of the problem, but some SDQ items are rated so that 0 = presence of the problem.
3. To measure favorable characteristics, the ASEBA has a variety of items for assessing competencies, adaptive functioning, and positive qualities, whereas the SDQ has prosocial items that are all rated in the same 0-1-2 format as the problem items, with which the prosocial and other favorably phrased items are interspersed.
4. The ASEBA obtains open-ended responses to clarify some problem items, to add and rate problems not already listed on the ASEBA forms, and to provide specific information about the child's activities, illnesses, and disabilities, the respondent's concerns about the child, and the best things about the child. The SDQ has an Impact Supplement that can be used to rate the child's overall distress and impairment in different areas, as well as the burden that the child's problems place on the family.
5. The ASEBA scales are scored on profiles in relation to norms appropriate for the child's gender, age, and cultural group, as well as for the type of informant. (We use 'informants' to include children who provide information about themselves.) Information about SDQ scores found for samples from various populations are available in published articles and at the website sdqinfo.com.
6. Hand-scored profiles, compact disk software, and internet scoring are available for the ASEBA, whereas the SDQ can be hand-scored with a

transparency. In addition, <http://www.sdqscore.net> gives free access and download to electronic scoring and produces a report online.

7. The ASEBA provides cross-informant comparisons in terms of side-by-side displays of problem item scores and scale scores, plus *Q* correlations between ratings by specific pairs of informants. (*Q* correlations measure the degree of agreement between ratings by each pair of informants.) Algorithms for combining data from multiple SDQ informants have been cited in published articles that are available at <http://www.sdqinfo.com>. The SDQ is also available free online at <http://www.sdqinfo.com> to any parent, teacher or young person (Goodman, Ford, Corbin, & Meltzer, 2004; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000; Goodman, Renfrew, & Mullick, 2000).

We turn now to reviews of the ASEBA and SDQ.

ASEBA instruments

The following sections outline the development of the ASEBA parent, teacher, and self-report forms, including their origins, scale development, and content. In subsequent sections, we present data on psychometrics and validity. Thereafter, we present multicultural findings.

Development of the ASEBA instruments

The earliest versions of the ASEBA instruments (Achenbach, 1966, 1978; Achenbach & Lewis, 1971) were developed to determine whether more syndromes of child psychopathology could be empirically identified than were included in the first and second editions of the American Psychiatric Association's (1952, 1968) *Diagnostic and Statistical Manual* (DSM-I and DSM-II). Based on reviews of the clinical research literature and consultation with clinicians, the initial items were formulated to describe a broad spectrum of problems that could be reported by parents, teachers, children, and mental health professionals. Items were added and revised on the basis of findings in psychiatric case records and feedback from different kinds of informants. Factor analyses of problems reported in child psychiatric case records and of problems rated by parents indeed revealed more syndromes than were included in DSM-I and DSM-II.

Parents' ratings were obtained on an early version of the Child Behavior Checklist (CBCL) completed for children who were assessed in a variety of mental health settings. On that version and subsequent versions of the CBCL, the Teacher's Report Form (TRF), and the Youth Self-Report (YSR), problem items have been rated *0 = not true (as far as you know)*, *1 = somewhat or sometimes true*, or *2 = very true or often true*. The standard rating period is 6 months for the CBCL and YSR and 2 months for

the TRF. These periods can be shortened if reassessments are to be done over shorter intervals.

The initial syndromes for scoring the CBCL, TRF, and YSR were derived from exploratory factor analyses (EFA) of CBCL, TRF, and YSR ratings for large clinical samples, analyzed separately for children of each gender, in different age ranges. The syndromes and competence scales were displayed on hand-scored and computer-scored profiles in relation to gender, age, and informant-specific norms obtained from parent, teacher, and self-ratings of large regional general population samples in the US (Achenbach & Edelbrock, 1983, 1986, 1987).

1991 Scales. EFAs of greatly enlarged clinical samples identified eight cross-informant syndromes that were common to both genders, different ages, and parent, teacher, and self-ratings (Achenbach, 1991a,b,c). The 1991 syndrome and competence scales were displayed on hand-scored and computer-scored profiles in relation to gender, age, and informant-specific norms based on a new US national probability sample.

Cross-informant comparisons and correlations. The 1991 computer-scoring program enabled users to print side-by-side displays of problem item and scale scores from parent, teacher, and self-ratings of each child and to display *Q* correlations between ratings by each pair of informants in relation to *Q* correlations for large reference samples.

The cross-informant comparisons and correlations were added because meta-analyses of findings with many instruments had yielded a mean correlation of only .28 between pairs of adults who play different roles vis-à-vis the children, including parents versus teachers versus mental health workers versus observers (Achenbach, McConaughy, & Howell, 1987). Between children's self-reports and reports by adults, the mean correlation was only .22. Even the mean correlation of .60 between pairs of adults who play similar roles (pairs of parents, teachers, mental health workers, observers) was not high enough for one informant's reports to substitute for reports by others playing similar roles. Although the meta-analyses were based on studies published before 1987, findings of modest cross-informant correlations have remained so consistent with the 1987 meta-analytic findings that they have been cited as being among 'the most robust findings in child clinical research' (De Los Reyes & Kazdin, 2005, p. 483). Modest cross-informant correlations are by no means limited to ratings of children, as meta-analyses have yielded a mean correlation of only .45 between self-ratings and collateral ratings of adult psychopathology (Achenbach, Krukowski, Dumenci, & Ivanova, 2005).

Agreement between different informants may be limited because the informants see children in different contexts, interact differently with the children, and have different mindsets for judging

and reporting on the children. Because each informant may reliably and validly report different aspects of children's functioning, disagreements between informants do not necessarily indicate errors. In fact, it has been found that reports by mothers, fathers, teachers, and children validly capture different genetic influences on children's problems (Arsenault et al., 2003; van der Valk et al., 2001). The value of obtaining data from multiple informants and the typically modest agreement between different informants make it essential to systematically obtain and compare reports by multiple informants for clinical evaluations of individual children and for research on groups of children.

2001 Scales. In 2001, slightly revised versions of the CBCL (now designated as the CBCL/6–18), TRF, and YSR were published (Achenbach & Rescorla, 2001). All three forms are scored on versions of the eight syndromes resulting from EFAs and confirmatory factor analyses (CFAs) of new samples of CBCLs, TRFs, and YSRs that included high-scoring nonreferred as well as clinically referred children. The syndromes are designated as *Anxious/Depressed*, *Withdrawn/Depressed*, *Somatic Complaints*, *Social Problems*, *Thought Problems*, *Attention Problems*, *Rule-Breaking Behavior*, and *Aggressive Behavior*. Second-order factor analyses yielded groupings of syndrome scales designated as *Internalizing* (the Anxious/Depressed, Withdrawn/Depressed, and Somatic Complaints syndromes) and *Externalizing* (the Rule-Breaking Behavior and Aggressive Behavior syndromes).

Another 2001 innovation was the addition of six DSM-oriented scales comprising items identified by experts from 16 cultures as being very consistent with DSM-IV diagnostic categories. The DSM-oriented scales are designated as Affective Problems, Anxiety Problems, Somatic Problems, Attention Deficit/Hyperactivity Problems, Oppositional Defiant Problems, and Conduct Problems. Empirically based subscales for Inattention and Hyperactivity-Impulsivity are based on factor analyses of the TRF Attention Problems syndrome, while DSM-oriented subscales are based on the international experts' judgments of items of the DSM-oriented Attention Deficit/Hyperactivity Problems scale. The 2001 CBCL and YSR are also scored on competence scales for activities, social relations, school, and total competence, while the TRF is scored on adaptive functioning scales for academic performance and favorable characteristics.

2007 Multicultural options and scales. In 2007, options for using multicultural norms were added to the computer program for the CBCL, TRF, and YSR (Achenbach & Rescorla, 2007a). The multicultural options enable users to display problem scale scores in relation to sets of norms based on societies that were found to have relatively low scores, medium

scores, or high scores. A child's CBCL, TRF, and YSR scale scores can be displayed in relation to different sets of norms. This can be especially helpful for evaluating immigrant children in relation to norms for both their home society and the host society. Additional innovations include scales designated as Obsessive-Compulsive Problems, Posttraumatic Stress Problems, Sluggish Cognitive Tempo, and Positive Qualities (YSR only).

ASEBA psychometric properties

Extensive psychometric data have been published for each edition of the ASEBA scales in many samples from many cultures. Table 2 summarizes psychometric findings obtained for the current editions of the scales in US samples (Achenbach & Rescorla, 2001).

As Table 2 shows, the internal consistencies (Cronbach's, 1951, alpha) were substantial for all sets of scales in the US samples. Averaged across the CBCL, TRF, and YSR, the mean US alphas were .96 for Total Problems, .92 for Internalizing and Externalizing, .82 for syndromes, and .81 for DSM-oriented scales. Psychometric findings from other societies have generally approximated those from the US. For example, averaged over the CBCL, TRF, and YSR samples from the 33 societies on which the 2007 multicultural norms were based ($N = 113,671$), the mean alphas were .94 for Total Problems, .87 for Internalizing and Externalizing, .76 for syndromes

Table 2 Psychometric properties of ASEBA scales

Scales	Alpha ^a	Test-retest reliability ^b	Long-term stability ^c
Empirically based syndromes			
CBCL/6–18	.83	.89	.70
TRF	.85	.89	.68
YSR	.79	.79	.52
DSM-oriented scales			
CBCL/6–18	.82	.88	.65
TRF	.84	.85	.65
YSR	.76	.79	.51
Internalizing and Externalizing			
CBCL/6–18	.92	.92	.81
TRF	.93	.88	.79
YSR	.90	.85	.56
Total Problems			
CBCL/6–18	.97	.94	.81
TRF	.97	.95	.77
YSR	.95	.87	.58

Note. Samples included US children referred for mental health services and nonreferred children, as detailed by Achenbach and Rescorla (2001). Except for 'Total Problems,' coefficients are means computed by averaging coefficients for all the relevant scales. Test-retest and cross-informant coefficients are mean r s computed by z transformation. All r s were $p < .05$.
^aAlphas for 3,210 CBCLs, 3,086 TRFs, and 1,938 YSRs.

^bMean test-retest intervals = 8 days for 73 CBCLs, 16 days for 44 TRFs, and 8 days for 89 YSRs.

^cMean stability intervals = 12 months for 75 CBCLs, 7 months for 144 YSRs, and 2 months for 22 TRFs.

and .74 for DSM-oriented scales (Achenbach & Rescorla, 2007a). Higher alphas have been found in clinical samples, such as those reported for German children by Döpfner et al. (1994, 1995a, b, 1997).

Although other kinds of psychometric data are not available from as many societies, good test-retest reliabilities have been found. For example, over intervals averaging 19 days, Leung et al. (2006) found a mean test-retest intraclass correlation of .85 for Total Problems, averaged across CBCLs, TRFs, and YSRs completed for a Hong Kong sample, compared to .92 for the U.S. sample over 8- to 16-day periods.

ASEBA validity findings

For instruments like the ASEBA and SDQ, three broad kinds of validity are relevant, as addressed in the following sections.

Content validity of ASEBA problem items. The most basic kind of validity is *content validity*, which refers to whether an instrument's items represent what the instrument is intended to assess. The problem items of the CBCL, TRF, and YSR were formulated to tap a broad spectrum of problems (a) that can be spontaneously reported by parents, teachers, and children with a minimum of inference and no need for highly trained interviewers, and (b) that discriminate significantly between children considered to need mental health and related services versus demographically similar children who are not considered to need such services.

The procedures for formulating, testing, and refining the items on the basis of data from many sources and feedback from many parents, teachers, children, and mental health professionals ensured that they describe problems that are of concern and can be understood by the intended respondents. Furthermore, ratings of individual problem items have been found to discriminate significantly between children referred for mental health services and demographically similar nonreferred children in several societies (e.g., Achenbach & Rescorla, 2001; Bilenberg, 1999; Fombonne, 1992; Montenegro, 1983; Verhulst, Prince, Vervuurt-Poot, & de Jong, 1989). The content validity of particular ASEBA items in relation to DSM-IV has also been supported by international experts' judgments that the items of the DSM-oriented scales are very consistent with DSM-IV diagnostic categories (Achenbach & Rescorla, 2001).

Criterion-related validity of ASEBA problem scales. *Criterion-related validity* refers to whether a particular measure agrees with external criteria that are more direct indicators of the target characteristics. Many kinds of analyses have supported the criterion-related validity of ASEBA scales in several societies. For example, in analyses of covariance,

multiple regressions, and other kinds of analyses, scores on the syndromes, DSM-oriented scales, Internalizing, Externalizing, and Total Problems have been significantly higher for clinically referred than nonreferred children, after controlling for demographic variables such as age, gender, SES, and ethnicity in US samples of demographically matched referred and nonreferred children, with *N*s ranging from 1,059 to 4,220 (Achenbach, 1991a,b,c; Achenbach & Rescorla, 2001, 2007a). Similar findings have been obtained in other societies, such as Denmark (Bilenberg, 1999), Finland (Helstelä, Sourander, & Bergroth, 2001), Chile (Montenegro, 1983), Germany (Schmeck et al., 2001), and the Netherlands (Verhulst, Akkerhuis, & Althaus, 1985; Verhulst et al., 1989). Although some of the demographic variables were significantly associated with the ASEBA scale scores, the effect sizes (ESs) for demographic variables were much smaller than the ESs for referral status.

Categorical analyses have been done to test the criterion-related validity of ASEBA scale scores that are in the normal range versus combined borderline and clinical ranges for discriminating between referred and nonreferred children. Odds ratios and chi squares have shown that significantly more referred than nonreferred children obtained scores in the borderline and clinical range on all ASEBA problem scales (Achenbach & Rescorla, 2001, 2007a). In a different type of categorical analysis, discriminant functions were computed to test the ability of ASEBA scale scores to correctly classify children as being referred versus nonreferred. After cross-validated correction for shrinkage via 'hold-one-out' procedures, the CBCL, TRF, and YSR scales correctly classified large percentages of children as referred versus nonreferred (Achenbach & Rescorla, 2001). Significant point-biserial correlations have also been found between DSM-IV clinical diagnoses and scores on the DSM-oriented scales (Achenbach & Rescorla, 2001).

Construct validity of ASEBA scales. 'Constructs' are mental abstractions that imply something more than single observations or scores. A dictionary definition of a construct is 'an object of thought constituted by the ordering or systematic uniting of experiential elements' (Gove, 1971, p. 489). Even when defined in theoretical terms, constructs should be anchored in data by being initially derived from data and by guiding the collection of new data that test the constructs while also advancing knowledge.

The study of child psychopathology has a rich history of theoretical constructs, such as libido, id, ego, superego, psychosexual stages, defense mechanisms, refrigerator parents who purportedly cause autism, and schizophrenogenic mothers. These constructs were to some extent based on clinicians' observations. However, they did not involve much systematic collection and organization of data for

purposes either of formulating the constructs or of testing them.

The kind of rich clinical theorizing that generated so many complex constructs has been largely replaced by a focus on constructs embodied in the DSM and ICD. Some constructs of adult disorders, such as schizophrenia and bipolar disorders, have been major foci of training, research, and services for over a century. Constructs of childhood disorders, by contrast, have much shorter histories and are still very much in flux. A major task is therefore to develop taxonomic constructs that are derived from data on child psychopathology and that also generate research to test the constructs while advancing knowledge. The syndromes derived from the CBCL, TRF, and YSR are designed to provide the basis for taxonomic constructs across the spectrum of behavioral, emotional, and social problems assessed by these instruments. As an exhaustive review of findings is beyond the scope of this article, we summarize several kinds of evidence supporting the ASEBA syndromes as valid measures of taxonomic constructs for children's behavioral, emotional, and social problems.

One kind of evidence for construct validity is agreement between a particular assessment procedure and other procedures for assessing similar constructs. Even if their conceptual basis differs, high correlations between different assessment procedures mean that they measure similar phenomena. The Conners (1997) Parent and Teacher Rating Scales include scales similar to the ASEBA Attention Problems and Aggressive Behavior syndromes and to the ASEBA DSM-oriented Attention Deficit/Hyperactivity Problems and Oppositional Defiant Problems scales. Correlations from .71 to .89 (mean = .81) have been found between corresponding scales of the ASEBA and Conners instruments for clinically referred children rated by parents and teachers (Achenbach & Rescorla, 2001). Correlations averaging .69 have been found between corresponding scales of the ASEBA and the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) completed separately by mothers, fathers, and teachers in a different clinical sample (Achenbach & Rescorla, 2001).

Evidence that is particularly relevant to the taxonomic construct validity of ASEBA syndromes has been provided by CFAs of CBCLs from 30 societies, TRFs from 20 societies, and YSRs from 23 societies (Total $N = 118,324$; Ivanova et al., 2007a, b, c). The fit of the data to the ASEBA syndrome structure was tested by computing the Root Mean Square of Approximation (RMSEA; Browne & Cudeck, 1993) for the CFAs that were done in the samples from each society. The RMSEA has been recommended as the best measure of goodness of fit for the kind of CFAs that were done (Yu & Muthén, 2002), with RMSEAs $< .06$ indicating good fit and RMSEAs between .06

and .08 indicating acceptable fit (Browne & Cudeck, 1993).

The mean of the RMSEAs in individual societies ranged from .032 for China to .054 for Portugal, indicated good fit for all societies except Lebanon, where the RMSEA of .063 qualifies as acceptable fit (Achenbach & Rescorla, 2007a). The CFA findings thus supported the taxonomic construct validity of the ASEBA syndromes across very diverse societies, although the data might also fit different models. Other kinds of factor analytic studies have generally supported earlier versions of some of the ASEBA syndromes, as well (e.g., in Israel, Auerbach & Lerner, 1991; the Netherlands, DeGroot, Koot, & Verhulst, 1994, 1996; Australia, Heubeck, 2000; China, Liu et al., 2000). An exception was a study by Hartman et al. (1999), who concluded that the 1991 CBCL and TRF syndrome structures were not supported by their analyses of data from several societies. Unfortunately, Hartman et al. included items that were too rarely endorsed and were too badly skewed for the type of analyses that they conducted. Nevertheless, the RMSEAs reported by Hartman et al. actually did support the 1991 ASEBA syndrome structure.

An important contribution of taxonomic constructs is to provide phenotypic markers for testing differential etiologies, developmental courses, and outcomes. Genetic studies in multiple societies have shown that ASEBA syndromes provide valid markers for genetic differences. As an example, genetic studies of British, Dutch, Swedish, and US twin samples have yielded moderate to high estimates of heritability for the ASEBA Aggressive Behavior syndrome (Arsenault et al., 2003; Edelbrock, Rende, Plomin, & Thompson, 1995; Eley, Lichtenstein, & Stevenson, 1999; Ghodsian-Carpey & Baker, 1987; van den Oord, Verhulst, & Boomsma, 1996). Furthermore, research shows that the developmental continuity of the Aggressive Behavior syndrome is influenced mainly by genetic factors, whereas the developmental continuity of the Rule-Breaking (previously called 'Delinquent') Behavior syndrome is affected more by shared environmental factors (Eley, Lichtenstein, & Moffitt, 2003). The more heritable developmental continuity of the Aggressive Behavior than the Rule-Breaking Behavior syndrome is consistent with the higher heritability found for the Aggressive Behavior than the Rule-Breaking Behavior syndrome in cross-sectional studies (Arsenault et al., 2003; Edelbrock et al., 1995; Eley et al., 1999; van den Oord et al., 1996).

The higher heritability of the Aggressive Behavior syndrome argues for making a taxonomic distinction between overtly aggressive behaviors and non-aggressive rule-breaking behaviors that are treated interchangeably in the diagnostic category of Conduct Disorder (CD). Another kind of evidence supporting the distinction between the taxonomic constructs embodied in the Aggressive Behavior and

Rule-Breaking Behavior syndromes consists of large correlations of the Aggressive Behavior syndrome (but not the Rule-Breaking syndrome) with low serotonergic activity and with beta-hydroxylase (DBH) levels (Birmaher et al., 1990; Gabel, Stadler, Bjorn, Shindlecker, & Bowden, 1993; Hanna, Yuweiler, & Coates, 1995; Stoff, Pollock, Vitiello, Behar, & Bridges, 1987). Equally important, longitudinal correlations for Aggressive Behavior syndrome scores at different ages have been significantly higher than for Rule-Breaking syndrome scores over multiple developmental periods (Stanger, Achenbach, & Verhulst, 1997).

In addition to the Aggressive Behavior and Rule-Breaking Behavior syndromes, the taxonomic construct validity of other ASEBA syndromes has been supported by genetic, developmental, and long-term outcome findings. For example, Attention Problems syndrome scores analyzed separately for Norwegian twin pairs of each gender at ages 5–9 and 12–15 years yielded heritabilities of 73% to 79% across the four gender/age groups (Gjone, Stevenson, & Sundet, 1996). In a longitudinal sample of Dutch twins, heritabilities for Attention Problems ranged from 70% to 74% when analyzed separately for each gender at different ages (Rietveld, Hudziak, Bartels, Van Beijsterveldt, & Boomsma, 2004). Furthermore, the developmental stability of Attention Problems scores was accounted for largely by genetic factors.

In a long-term test of the predictive power and outcomes of ASEBA syndromes, Hofstra, van der Ende, and Verhulst (2002) found that ASEBA syndromes differentially predicted adult DSM-IV diagnoses 14 years later in a Dutch probability sample. Numerous other long-term outcome findings have been reported from this study as well as from clinical and nonclinical samples in the Netherlands, Australia, and the US (e.g., Heijmens Visser, van der Ende, Koot, & Verhulst, 2000; MacDonald & Achenbach, 1999; Sawyer, Mudge, Carty, Baghurst, & McMichael, 1996; Stanger, MacDonald, McConaughy, & Achenbach, 1996). There is thus abundant support for the taxonomic construct validity of ASEBA syndromes. However, much remains to be done by way of testing many other kinds of distinctions among the taxa represented by the ASEBA syndromes.

Multicultural ASEBA findings

Many ASEBA findings cited so far have counterparts in more than one society. Although many studies were done in single societies, CFA support for the 2001 ASEBA syndromes was obtained in uniform CFA analyses of CBCL samples from 30 societies, TRF samples from 20 societies, and YSR samples from 23 societies (Ivanova et al., 2007a, b, c). In this section, we summarize studies that have applied multicultural dimensional methodology to large samples from multiple societies.

Scale score differences among multiple societies. Following a series of studies that compared ASEBA scores and correlates in pairs of societies (e.g., Achenbach, Verhulst, Edelbrock, Baron, & Akkerhuis, 1987; Achenbach et al., 1990), the first published comparisons of multiple societies included 13,697 CBCLs from 12 societies (Crijnen, Achenbach, & Verhulst, 1997). For Internalizing, Externalizing, and Total Problems scores, ESs for differences among societies ranged from 7% to 11%, which are medium ESs according to Cohen's (1988) criteria. Crijnen et al. (1999) subsequently compared scores obtained on the eight 1991 syndromes in the nine societies that had CBCL data for ages 6–17 ($N = 11,887$). ESs for differences among societies were small (1% to 5%) for five syndromes and medium (6% to 9%) for three syndromes. A similar multicultural comparison of 1991 YSR problem scale scores yielded ESs of 3% to 8% for differences among seven societies ($N = 7,137$; Verhulst et al., 2003).

The procedures pioneered by Crijnen et al. (1997) (1999) were subsequently applied to 2001 ASEBA scales scored from CBCLs completed in 31 societies, TRFs in 21 societies, and YSRs in 24 societies (total $N = 113,671$; Rescorla et al., 2007a, b, c). The societies were in Africa, Asia, Australia, the Caribbean, Europe, the Middle East, and North America. Like the Crijnen and Verhulst studies, the Rescorla studies found small to medium ESs for differences among societies, as well as gender and age effects that were quite consistent across societies. Differences between societies also had small to medium ESs on the 2001 DSM-oriented scales, which had not been available for the Crijnen and Verhulst studies.

The Rescorla et al. studies make it possible to identify scales whose scores varied least versus most among the different societies. The DSM-oriented Conduct Problems scale showed the smallest ESs for differences among societies on the CBCL (4% ES, averaged over separate analyses for ages 6 to 11 and 12 to 16), TRF (3% ES for ages 6 to 15), and YSR (3% ES for ages 11 to 16). Among the problem scales, the Anxious/Depressed syndrome showed the largest ES on the CBCL (11%) and TRF (13%), whereas the Attention Problems syndromes showed the largest ES on the YSR (9%). The ESs for differences among societies on the Internalizing scale exceeded those for Externalizing (10% vs. 6% on the CBCL, 11% vs. 4% on the TRF, and 6% vs. 5% on the YSR). Parent, teacher, and self-ratings of conduct problems thus tended to be quite similar across societies, whereas parent and teacher ratings of Internalizing problems, especially the Anxious/Depressed syndrome, varied more across societies. However, on the YSR Positive Qualities scale, societal differences had a much larger ES (27%) than was found for any of the problem scales. The 27% ES reflected much less within-society variance than was found on the problem scales. In other words, self-ratings of favorable characteristics were much more uniform within each

society than were parent, teacher, or self-ratings of problems.

Gender-by-age interactions. Certain scale scores showed gender-by-age effects that were very consistent despite major differences in ethnicity, language, culture, political/economic system, religion, level of economic development, and region of the world. For example, girls' tendency to obtain higher CBCL Internalizing scores than boys was relatively small at ages 6–8 years but, because girls' Internalizing scores increased with age much more than boys' scores did, the gender difference became much larger by ages 15–16. Conversely, boys' CBCL Aggressive Behavior scores were much higher than girls' scores at ages 6–8, but declined more with age than girls' scores did until the gender difference became negligible at ages 15–16.

Informant differences. The availability of CBCL, TRF, and YSR data from so many societies makes it possible to identify similarities and differences between findings obtained from the different kinds of informants. For example, in the 19 societies for which CBCL and YSR ratings were available for the same youths, the YSR scores were significantly higher than the CBCL scores for problem items that had counterparts on both forms (Rescorla et al., 2007c). Across many very different societies, youths thus tend to report more problems for themselves than their parents do.

Another interesting finding was that parents and teachers in nearly every society rated boys higher than girls on the DSM-oriented Attention Deficit/Hyperactivity problems scale, but YSR self-ratings on this scale did not show a significant gender difference. The importance of multi-informant assessment was underlined by other findings, as well. For example, unlike parents and youths, teachers did not consistently rate girls higher than boys on any problem scales. Furthermore, gender differences on the Attention Problems, Rule-Breaking Behavior, and Aggressive Behavior syndromes (boys higher) were larger in teacher ratings than in parent ratings.

For cross-informant agreement about individual children, the correlations of CBCL with TRF scores, CBCL with YSR scores, and TRF with YSR scores, averaged over all societies that had these combinations of informants, were generally similar to the cross-informant correlations for US samples (Achenbach & Rescorla, 2001, 2007a).

Item scores. In addition to comparisons of scale scores and their correlates, the data obtained with the same instruments in so many societies make it possible to test the degree to which people in different societies rate the same problem items as relatively high, medium, or low. Rescorla et al. (2007a, b, c) did this by computing correlations between the mean score obtained on each problem item in each

society and the mean score on each item in each other society. In other words, a correlation was computed between the CBCL mean item scores in Society 1 and Society 2, Society 1 and Society 3, and so on for all pairs of the 31 societies from which CBCL data were available. The same was done for each pair of societies that had TRF data and for each pair of societies that had YSR data. To summarize the level of cross-society agreement, Rescorla et al. then averaged the bi-society correlations obtained for every pair of societies. For both the CBCL and TRF, the mean of the bi-society correlations was .74, while for the YSR it was .69. All three correlations are large according to Cohen's (Cohen, 1988) criteria, indicating substantial multicultural similarity in the items that received relatively high, medium, or low ratings. Further, Roessner et al. (2007) detected similar CBCL profiles of hyperactive children from Brazil and Germany.

SDQ instruments

The following sections regarding the parent, teacher, and self-report versions of the SDQ are organized along lines like those for the ASEBA. However, owing to the design of the SDQ instruments for screening based on the Rutter (1967) parent questionnaire, their lack of DSM-oriented scales, and their shorter history, there are fewer findings than for the ASEBA.

Development of the SDQ instruments

The published history of the SDQ instruments began with Robert Goodman's (1994) report of having parents of 320 Greater London children with hemiplegia complete a modified version of a questionnaire that had originally been developed by Sir Michael Rutter (1967; Rutter, Tizard, & Whitmore, 1970). Goodman's questionnaire included the 31 items of the Rutter questionnaire, plus 5 new problem items and 14 items concerning 'desirable traits.' Parents rated each item as *doesn't apply*, *applies somewhat*, and *certainly applies*. From the results of a principal components analysis (PCA) with varimax rotation of the parents' ratings, Goodman concluded that a 'solution with just six factors was the easiest to interpret and made the greatest clinical sense' (p. 1491). He designated the factors as *hyperactivity/inattention*, *prosocial behavior*, *conduct problems/oppositionality*, *somatic/developmental*, *internalization*, and *peer relationships*.

To construct the SDQ, Goodman (1997) wrote five items to form scales for each of five dimensions suggested by his PCA of the modified Rutter questionnaire. (He did not write items for the somatic/developmental dimension.) The SDQ scales were designated as *Hyperactivity*, *Emotional Symptoms*, *Conduct Problems*, *Peer Problems*, and *Prosocial Behavior*. SDQ items are rated *not true*, *somewhat true*, or *certainly true*, with scores of 0-1-2 being

given to items that describe unfavorably phrased problem items and 2-1-0 to prosocial items and favorably phrased problem items. Parents of 386 children attending London psychiatric or dental clinics and teachers of 185 of the children completed the SDQ and Rutter questionnaires. For the Conduct Problems, Emotional Symptoms, and Hyperactivity scales that had counterparts on both the SDQ and Rutter questionnaires, correlations with the Rutter scales ranged from .78 to .91 (mean = .83 for parents' ratings, .89 for teachers' ratings). SDQ and Rutter Total Deviance/Difficulties scores correlated .88 for parents' ratings and .92 for teachers' ratings. These high correlations indicate very similar rank orders of scores obtained on corresponding scales of the SDQ and Rutter questionnaires.

Because Goodman (1997) wrote the SDQ items and assigned them to scales suggested by his 1994 factor analyses of the modified Rutter questionnaire, the SDQ scales themselves were not actually constructed by bottom-up factor analytic or other statistical procedures. However, Goodman (2001) subsequently reported factor analyses of SDQs for a British general population sample of 9,998 5- to 15-year-olds rated by parents, 7,313 5- to 15-year-olds rated by teachers, and 3,983 11- to 15-year-olds who rated themselves. The Goodman (2001) article did not specify the measure of association between items or the method for extracting factors. However, Table 2 in the Goodman article indicates that factor loadings were obtained from a varimax rotation of five factors. According to Table 2, the factor loadings were consistent with Goodman's assignment of items to the SDQ scales, except that two items rated by teachers and two rated by youths loaded on scales other than their assigned scales at levels approximating or exceeding their loadings on their assigned scales.

SDQ psychometric properties

Goodman (2001) reported alphas and 4–6-month stabilities of SDQ scale scores in the British general population samples for which he reported factor loadings. Table 3 summarizes these alphas and stabilities for Total Difficulties and for the mean of the four difficulties subscales scored from the parent, teacher, and youth-rated SDQs. Ranging from a mean of .59 for the youth SDQ to .78 for the teacher SDQ, the alphas for the difficulties subscales may be limited by their small numbers of items. The overall mean alpha of .83 for Total Difficulties was more satisfactory. The mean 4–6-month stability correlation was .64 for the difficulties subscales and .71 for Total Difficulties.

With respect to test–retest reliabilities over shorter periods when children's behavior is not expected to change much, Goodman (1999) reported an intra-class correlation of .85 for Total Difficulties scores on parent SDQs, but did not report results for the

Table 3 Psychometric properties of SDQ scales

Scales	Alpha ^a	Test–retest reliability ^b	Long-term stability ^c
Difficulties scales			
Parent SDQ	.66	.71	.64
Teacher SDQ	.78	.73	.73
Youth SDQ	.59	.71	.56
Total Difficulties			
Parent SDQ	.82	.81	.72
Teacher SDQ	.87	.74	.80
Youth SDQ	.80	.79	.62

Note. Except for Total Difficulties, coefficients are means computed by averaging the SDQ Emotional Symptoms, Conduct Problems, Hyperactivity-Inattention, and Peer Problems scales. Test–retest coefficients are mean *rs*. All *rs* were $p < .05$. ^aAlphas for 9,998 Parent SDQs, 7,313 Teacher SDQs, and 3,983 Youth SDQs in a British general population sample (Goodman, 2001).

^bTest–retest interval = 2 weeks for 120 parent, teacher, and youth SDQs in an Australian general population sample (Mellor, 2004).

^cMean stability intervals = 4 to 6 months for 2,091 Parent SDQs, 796 Teacher SDQs, and 781 Youth SDQs in a British general population sample (Goodman, 2001).

subscales. Mellor (2004) reported 2-week test–retest reliabilities for a general population sample of Australian children. As summarized in Table 3, the test–retest correlations for difficulties subscales ranged from .71 for parent and youth SDQs to .73 for teacher SDQs, with a mean of .72. For Total Difficulties, the test–retest correlations ranged from .74 for teacher SDQs to .81 for parent SDQs, with a mean of .78.

Psychometric data from other societies include mean alpha coefficients for the SDQ's specific difficulties scales that have ranged from a low of .54 for self-ratings in a Dutch general population sample to a high of .79 for teacher ratings in the same sample (Van Widenfelt, Goedhart, Treffers, & Goodman, 2003). Alpha coefficients for SDQ Total Difficulties have ranged from .70 for self-ratings to .88 for teacher ratings in the same study. Alphas were between these extremes in studies from Finland (Koskelainen, Sourander, & Kaljonen, 2000), Germany (Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004), Russia (Ruchkin, Kuposov, & Schwab-Stone, 2007), Sweden (Smedje, Broman, Hetta, & von Knorring, 1999), and a multi-European sample (Becker et al., 2006). The mean alphas for the four difficulties subscales thus ranged from the .50s to the .70s, while the alphas for Total Difficulties ranged from the .70s to the .80s in different societies.

SDQ validity findings

Content validity of SDQ problem items. Modeled on the Rutter (1967) parent and teacher questionnaires, the SDQ problem items were written by Goodman (1997) for Emotional Symptoms, Hyperactivity, and Conduct Problems scales to correspond to scales of the same names scored from the Rutter

questionnaires. The original versions of the SDQ items were 'modified and amalgamated on the basis of a succession of informal trials as well as advice from colleagues' (Goodman, 1997, p. 582). Correlations of .78 to .92 between the corresponding Rutter and SDQ scales scored from parent and teacher ratings indicated that the two sets of scales were tapping similar content (Goodman, 1997).

Criterion-related validity of SDQ problem scales. Goodman (1997) described the SDQ as a 'behavioral screening questionnaire' (p. 581). Consequently, a primary type of criterion-related validity would be agreement of SDQ scores with classification of children as having versus not having mental health problems on the basis of criteria external to the SDQ. In the first publication of findings on this kind of criterion-related validity, Goodman (1997) reported Receiver Operating Characteristics (ROC) analyses showing that Total Difficulties scores from both parent and teacher SDQs discriminated significantly between 244 children attending two psychiatric clinics versus 159 attending a dental clinic in London. Age, gender, SES, and ethnicity were not controlled, although it was stated that 'closely similar results were obtained when ROC and correlational analyses were repeated separately for boys and girls and for children aged 4–10 and 11–16' (p. 583). Total Problems scores on the Rutter parent and teacher questionnaires discriminated between children attending the psychiatric versus dental clinics at levels similar to those obtained with the SDQ.

In another study employing ROC analyses of SDQ scores for children attending London psychiatric clinics versus a dental clinic, SDQs completed by mothers yielded a mean area under the curve (AUC, a measure of the accuracy of classification) of .83, averaged across the difficulties scales (Goodman & Scott, 1999). A subset of CBCL scales deemed analogous to the SDQ scales yielded a mean AUC of .86, but there were no significant differences between the AUCs for the SDQ and CBCL. Although the authors stated that 'a similar pattern of results emerged when boys and girls were analyzed separately' (p. 21), their analyses did not control for age, gender, SES, or ethnicity.

In a pilot study for the 1999 British Child Mental Health Survey, Goodman (1999) compared SDQ scores for 232 children attending three London mental health clinics versus 467 in a general population sample. An Extended Version of the SDQ was used that included questions regarding the impact of the child's problems in terms of the magnitude of difficulties, their duration, the degree to which they distress the child and interfere with the child's home life, friendships, classroom learning, and leisure activities, and the degree of burden on the child's family. For cutoff scores selected to maximize associations with referral status in the sample that was analyzed, kappa coefficients averaged .55 between

referral and Total Difficulties scores in parent, teacher, and self-ratings. Impact scores computed in various ways yielded kappas averaging .61 to .64 in relation to referral status. Although referral status was significantly associated with Total Difficulties scores, these associations were significantly weaker than the associations between impact scores and referral status.

Studies outside the UK have also found significant associations between referral for mental health or special education services and various combinations of difficulty scores and impact scores (e.g., in the US, Bourdon, Goodman, Rae, Simpson, & Koretz, 2005; Germany, Klasen et al., 2000). However, like the Goodman studies in the UK, these studies failed to control for age, gender, SES, or ethnic differences, which could be associated with referral status in ways that would affect associations between SDQ scores and referral for services.

Construct validity of SDQ scales. Although the SDQ was developed primarily as a brief screening instrument, its four difficulties subscales can be viewed as representing broad constructs of hyperactivity, emotional symptoms, conduct problems, and peer problems. One way to evaluate the construct validity of these subscales is by testing their agreement with other measures of similar constructs. Accordingly, correlations between these SDQ subscales and ASEBA scales have been computed in several studies. In the previously cited Goodman and Scott (1999) study of children attending dental clinics or psychiatric clinics in London, mothers' ratings on the SDQ subscales correlated from .59 to .84 (mean = .72) with CBCL scales deemed to be similar. In a Finnish general population sample, corresponding subscales of SDQs and CBCLs completed by parents had correlations of .34 to .70, with a mean of .52. Between SDQs and YSRs completed by children, the correlations ranged from .43 to .68, with a mean of .60. In a Dutch general population sample, correlations ranged from .51 to .78 (mean = .68) between corresponding scales of SDQs and CBCLs completed by parents and from .41 to .66 (mean = .58) between SDQs and YSRs completed by children (Van Widenfelt et al., 2003). In a German study, correlations between SDQ and CBCL scales ranged from .60 to .76 (mean = .67) in a community sample and from .68 to .81 (mean = .73) in a clinical sample (Klasen et al., 2000). For self-ratings, correlations between SDQ and YSR scales ranged from .58 to .78 (mean = .67) in the community sample, but were not reported for the clinical sample. And in another German study of psychiatric inpatients and outpatients, SDQ scales correlated with CBCL scales from .64 to .82 (mean = .75) and with YSR scales from .64 to .82 (mean = .79; Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004).

Although no SDQ scales were formulated according to DSM or ICD diagnostic criteria, the designation of scales as Conduct Problems and Hyperactivity implies links to the diagnostic categories of Conduct Disorder and ADHD, respectively. Furthermore, Goodman and colleagues have developed a computerized algorithm for using SDQ scores to predict diagnoses aggregated into the following groups: Conduct and oppositional disorders; hyperactivity and inattention disorders; and anxiety, depressive, and obsessive-compulsive disorders (Goodman, Renfrew, & Mullick, 2000). The algorithm uses SDQ difficulties and impact scores in various combinations to generate 'unlikely,' 'possible,' and 'probable' ratings for each group of disorders. Developed on 101 4- to 16-year-olds who received ICD-10 diagnoses at a London mental health clinic, the algorithm was tested on 89 4- to 16-year-olds who received ICD-10 diagnoses at a mental health clinic in Dhaka, Bangladesh. For each child, a clinician rated diagnoses from the three broad groups as 'absent,' 'borderline,' or 'definite.' For each diagnostic category in each clinic, a 3×3 table was constructed for the relations between the SDQ predictions as unlikely, possible, or probable and the ratings of the diagnoses as absent, borderline, or definite. Kendall's tau b statistic for the associations between the 3-step SDQ predictions and the 3-step ratings of diagnoses ranged from .50 to .67 in the London clinic and from .49 to .73 in the Dhaka clinic, with an overall mean of .60.

Goodman's algorithm was also applied to 7,954 5- to 15-year-olds in the 1999 British Child Mental Health Survey for whom SDQs were completed and ICD-10 diagnoses were made with the DAWBA from interview and questionnaire data obtained from the same informants who completed the SDQs. Because there are marked discrepancies between the ICD-10 criteria for hyperkinetic disorders and the DSM-IV criteria for ADHD, findings were reported separately for diagnoses made according to each set of criteria. Although the SDQ predictions were again graded as unlikely, possible, or probable, they were dichotomized as unlikely and possible versus probable for analyses in relation to present versus absent diagnoses in terms of sensitivity (i.e., agreement between the SDQ and DAWBA that any psychiatric disorder was present). Specificity (i.e., agreement that a certain disorder was absent), which is the other side of the screening coin, was not reported for categories of disorders since the SDQ was primarily designed to identify children with high risk for psychiatric disorders who therefore warrant more careful assessment. The SDQ's sensitivity was 63.3% for identifying any disorder identified by the DAWBA, 76.2% for any conduct/oppositional disorder, 86.1% for any ICD-10 hyperkinetic disorder, and 75.4% for any DSM-IV ADHD disorder. Rather than reporting findings for any emotional disorder, Goodman et al. reported a sensitivity of 74.6% for

any depressive disorder and 50.5% for any anxiety disorder.

For the same broad categories as were used in the 1999 British Mental Health Survey, Goodman, Ford, Corbin, and Meltzer (2004) reported sensitivities of 82.7% to 97.7% for SDQ predictions of DAWBA diagnoses among 1,025 British children in residential or foster care. However, specificities were not reported for this sample, either. In a Norwegian sample, data combined from parent and teacher SDQs achieved a sensitivity of 77.3% and specificity of 85.1% for any diagnosis made from the DAWBA administered to the parents (Hysing, Elgen, Gillberg, Lie, & Lundervold, 2007).

A study of German inpatient and outpatient children used ROC analyses of relations between continuous scores on the SDQ Conduct Problems, Emotional Problems, and Hyperactivity scales and ICD-10 clinical diagnoses grouped into three broad categories corresponding to the SDQ scales (Becker, Hagenberg, Roessner, Woerner, & Rothenberger, 2004). For SDQ self-ratings, the AUC ranged from .684 to .773, with a mean of .716, while for SDQ parent ratings, the AUC ranged from .712 to .824, with a mean of .765.

Taken together, several studies indicate significant associations between SDQ scores and ICD-10 diagnoses made in various ways, although more meaningful evaluations could be made if specificities were presented along with sensitivities. Proper interpretation of sensitivities requires that specificities also be considered for the following reason: If Instrument A diagnoses every child as having a particular disorder, its sensitivity in relation to positive diagnoses made according to Criterion B would be 100%; unfortunately, this high level of sensitivity neglects the fact that Instrument A's failure to diagnose any child as free of the disorder would yield a specificity of 0%.

Genetic studies in different societies have shown that SDQ scales provide valid markers for genetic differences. Genetic studies of British and US twin samples have yielded moderate to high estimates of heritability for SDQ *Hyperactivity* and *Conduct Problems* scores (Martin, Scourfield, & McGuffin, 2002; Gregory, Eley, & Plomin, 2004; Scourfield, Van den Bree, Martin, & McGuffin, 2004). SDQ parent and teacher *Hyperactivity* scores analyzed separately for British twin pairs of each gender at ages 5–16 years yielded heritabilities of 70% to 81% across the age groups (Martin et al., 2002). A combination of SDQ parent, teacher and self-ratings also revealed a highly heritable phenotype of pervasive conduct problems.

Among factor analytic studies of the SDQ, one that involved children who probably differed the most from Goodman's British samples was done on teachers' SDQ ratings of 1,187 7–9-year-old students in Kinshasa, Democratic Republic of Congo (Kashala, Elgen, Sommerfelt, & Tylleskar, 2005).

In a PCA with varimax rotation of five factors, 16 of the 25 SDQ items loaded on factors corresponding to SDQ scales. In terms of each item's highest loadings, the factor designated by Kashala et al. as 'prosocial' had four items from the Prosocial scale, two items from the Peer Problems scale, and one item from the Conduct Problems scale. The factor designated as 'hyperactivity' had three items from the Hyperactivity scale and one item from the Prosocial scale. The factor designated as 'emotional symptoms' had five items from the Emotional Symptoms scale and one from the Peer Problems scale. The factor designated as 'conduct problems' had three items from the Conduct Problems scale and one from the Peer Problems scale. And the fifth factor had two items from the Conduct Problems scale and two from the Hyperactivity scale. There was thus no factor corresponding to the Peer Problems scale.

In one of several European studies, PCA-varimax analyses were applied to parents' SDQ ratings of 930 German 6–16-year-olds assessed in a national survey of weight problems and eating habits (Woerner, Becker, & Rothenberger, 2004). The results clearly supported the five SDQ scales in that the one loading over .40 found for each of the 25 items was on the factor corresponding to the appropriate SDQ scale. In another European PCA-varimax study, the data consisted of parents' SDQ ratings of 900 Swedish children from separate nonclinical samples of 6–8-year-olds and 10-year-olds (Smedje et al., 1999). Although there were substantial gender differences in loadings on a few items and a few loaded more heavily on factors other than the intended ones, Smedje et al. concluded that the results generally supported the SDQ scale structure. A similar conclusion was reached in a European study of 1,459 6–18-year-old children with ADHD (Becker et al., 2006).

EFA of SDQ self-ratings by 1,458 13–17-year-old students in two Finnish cities yielded somewhat different results for boys and girls (Koskelainen, Sourander, & Vauras, 2001). When a 5-factor solution was forced, factors corresponding to the SDQ Emotional Symptoms and Prosocial scales were obtained for both genders. A factor corresponding to the Conduct Problems scale was obtained only for girls, while a factor corresponding to the Hyperactivity scale was obtained only for boys. However, the remaining two factors found for each gender did not clearly match SDQ scales. When Koskelainen et al. redid the analyses without forcing a 5-factor solution, they found the following three factors that were very similar for both genders: a factor that included items from the SDQ Hyperactivity and Conduct Problems scales; a factor comprising prosocial items; and a factor that included items from the SDQ Emotional Symptoms and Peer Problems scales. PCAs of 1,111 Dutch students' SDQ self-ratings also failed to support

separate factors corresponding to the Peer and Conduct Problems scales (Muris, Meesters, Eijkelboom, & Vincken, 2004). A 4-factor solution yielded a factor comprising conduct and peer problems, while a prosocial factor included favorably worded items from multiple SDQ scales.

Other studies have used CFA to test the scale structure of the SDQ. For example, in a Norwegian study, CFA was used to test the structure of SDQ self-ratings by 4,167 11–16-year-old students (Rønning, Handegaard, Sourander, & Mørch, 2004). Of the four indices of fit that were used to evaluate the CFA, only the goodness-of-fit index (GFI) indicated good fit, with a value of .98. Values for the comparative fit index (CFI; .82), standardized root mean square of approximation (SRMR; .11), and RMSEA (.47) deviated considerably from limits for adequate fit. Based on further analyses, Rønning et al. suggested ways in which the model fit could be improved, such as by moving some favorably worded items from difficulties scales to the Prosocial scale. A CFA of students' self-rated SDQs from Arkhangelsk, Russia, yielded goodness of fit indices that supported the five SDQ scales (Ruchkin et al., 2007).

PCA, EFA, and CFA were all applied to parents' SDQ ratings of 9,574 4–17-year-olds assessed in the US National Health Interview Survey (Dickey & Blumberg, 2004). Ten SDQ items were reworded to make them more appropriate for American parents and for older children. PCA-varimax analyses like those used by Goodman supported three SDQ scales, but did not clearly support the Conduct or Peer Problems scales. EFA performed on a randomly selected half of the sample yielded three factors that were designated as prosocial problems (or 'positive construal' because the factor comprised most of the SDQ's favorably worded items), externalization problems, and internalization problems. When tested via CFA in the other half of the sample, the 3-factor model fit well, according to $GFI = .97$ and root-mean-square-residual (RMR) = .06, both of which were within the range for good fit. Substitution of Pearson correlations for polychoric correlations also yielded good fit for the 3-factor model, according to both $GFI = .94$ and $RMR = .04$.

Multicultural findings

Many of the studies cited so far have reported mean scores and other findings for SDQ scales in single societies. Some have also made descriptive comparisons with mean scores and/or particular cutpoints, such as the 90th percentile scores, obtained in Goodman's UK samples. An article by Obel et al. (2004) presents descriptive comparisons between mean SDQ scale scores for various samples from Norway, Denmark, Finland, and Sweden. However, rigorous statistical comparisons of scores from

multiple societies have evidently not been published. Because firm conclusions cannot be drawn about multicultural variations in problems, we will just summarize findings for mean scale scores for various groups.

The Obel et al. (2004) article is a useful starting point, because it provided mean SDQ parent ratings for one Danish, two Norwegian, and two Swedish samples of 7-year-olds, with *Ns* ranging from 290 to 4,968. The mean Total Difficulties scores ranged from 5.7 in Bergen, Norway, to 7.2 in Gävleborg, Sweden, with *SDs* ranging from 4.4 to 5.1. The difference of 1.5 between the largest and smallest mean was .32 of the mean *SD* of 4.7, which is a small ES (Cohen, 1988). Obel et al. also provided mean scores for SDQ self-ratings by 15-year-olds in one Finnish sample and three Norwegian samples. *Ns* ranged from 587 to 3,265, mean Total Difficulties scores ranged from 9.8 to 11.4, and *SDs* ranged from 4.9 to 5.8. The difference of 1.6 between the largest and smallest means was .29 of the mean *SD*, again indicating a small ES. However, the difference of 4.4 between the overall mean of 10.7 for self-ratings by 15-year-olds versus 6.3 for ratings by parents of 7-year-olds was .86 of the mean *SD*, a large ES that indicates a need for different norms and cutpoints for parent versus self-ratings of these different age groups.

In their report of parent SDQ ratings for 9,878 US 4–17-year-olds, Bourdon et al. (2005) compared their mean Total Difficulties score of 7.1 with the mean of 8.4 reported by Goodman (1997) for a British sample. Although no statistical tests were reported, descriptive comparisons between the percentages of children who would be classified by various cutpoints for low, medium, and high difficulties were consistent with the apparent tendency for US parents to rate their children lower on Total Difficulties than UK parents.

Van Widenfelt et al. (2003) compared SDQ scores for parent, teacher, and self-ratings of Dutch 11–16-year-olds with British SDQ scores posted at <http://www.sdqinfo.com>. Differences in mean scale scores exceeded .2 of the pooled *SD* for parent ratings of Total Difficulties, Conduct Problems, and Peer Problems, with Dutch scores being lower than British in all comparisons. According to the criterion of $>.2$ *SD*, Dutch self-ratings were also lower than British for Emotional Symptoms and Prosocial, but Dutch self-ratings were higher for Peer Problems and Dutch teacher ratings were higher for Peer Problems and Prosocial.

Numerous studies have reported significant gender differences in parent, teacher, and self-ratings on various SDQ scales in different societies (e.g., Koskelainen et al., 2000, 2001; Muris et al., 2004; Rønning et al., 2004; Woerner et al., 2004). However, multicultural analyses are needed to test the consistency of the gender effects for each scale, age group, and informant.

Integration of findings

Findings obtained by etic methods in diverse cultures can advance knowledge in ways analogous to meta-analytic findings from methodologically diverse studies. Just as meta-analytic findings deserve more confidence than findings from one study or a handful of methodologically similar studies, findings that are consistent across diverse cultures deserve more confidence than findings from one culture or a handful of similar cultures. Equally important, findings that are consistent in most but not all cultures and findings that differ markedly among most cultures invite emic investigations to identify reasons for the differences.

Feasibility of dimensionally based assessment in diverse contexts

Our reviews of ASEBA and SDQ research have documented the feasibility of using dimensionally based assessment in very diverse cultural contexts. The basic finding that hundreds of thousands of parents, teachers, and children from the diverse cultural groups listed in Table 1 have been willing and able to complete ASEBA and SDQ forms supports the applicability of this approach to multicultural assessment. Although cultural groups may yet be found where parents, teachers, and/or children are unwilling or unable to provide standardized ratings of children's problems, the findings to date indicate that people from remarkably diverse backgrounds understand the format and content of these forms. Because lay interviewers can quickly administer ASEBA and SDQ forms, dimensionally based assessment can obtain data at low cost from people who cannot complete forms independently.

Issues of translation and relevance

It must, of course, be recognized that translations may not achieve precisely equivalent meanings for all items in all languages and cultural groups. It must also be recognized that (a) certain items on standardized forms may not be equally relevant to all cultural groups and (b) standardized forms may fail to tap problems that are relevant to certain groups. Nevertheless, the following kinds of evidence indicate that translational imprecision and the differential relevance of particular items have not been major obstacles to obtaining comparable data for many cultural groups:

1. Despite statistically significant differences between mean problem scores in epidemiological samples from different populations, the mean scores span relatively narrow ranges across very diverse populations (Obel et al., 2004; Rescorla et al., 2007a, b, c).

2. The mean scores on individual problem items obtained in each population correlate highly with the mean scores obtained on these items in most other populations (Crijnen et al., 1997; Verhulst et al., 2003; Rescorla et al., 2007a, b, c).
3. Age and gender effects on problem scores are similar in most populations (Rescorla et al., 2007a, b, c).
4. In 16 societies for which relevant data have been published, problem scores have been negatively associated with SES (reviewed by Achenbach & Rescorla, 2007b).
5. Internal consistencies and test-retest reliabilities of scale scores have been substantial in the populations for which psychometric findings have been published (e.g., Achenbach & Rescorla, 2001, 2007a; Goodman, 2001; Leung et al., 2006; Mellor, 2004; Rescorla et al., 2007a, b, c).
6. Uniform CFA procedures have supported the empirically based ASEBA syndromes in many populations (Ivanova et al., 2007a, b, c). Various factor-analytic procedures have produced mixed findings for the a priori SDQ scales in several populations (e.g., Dickey & Blumberg, 2004; Goodman, 2001; Kashala et al., 2005; Koskelainen et al., 2001; Rønning et al., 2004; Woerner et al., 2004).

Distributions of scale scores

Another important finding is that distributions of problem scale scores overlap considerably among the populations in which epidemiological samples have been obtained. Despite the fact that significantly higher mean problem scale scores were found in some populations than others, the large overlaps between the distributions of scores show that many children in each population obtained scores as high (or low) as many children in each other population. The range of scores and the standard deviations within each epidemiological sample were considerably larger than the range and the standard deviation of mean scores across all populations analyzed (Rescorla, 2007a, b, c). In other words, the differences within populations were larger than the differences between populations.

Just as each population's children are diverse with respect to their genotypic characteristics, so, too, dimensional assessment instruments show that they are diverse with respect to the phenotypic behavioral/emotional/social problems reported for them. Such findings argue against tendencies to view cultures as 'internally homogeneous and externally distinctive' (Hermans & Kempen, 1998) with respect to parent, teacher, and self-reports of children's problems. On the contrary, the findings indicate that internal heterogeneity exceeds external distinctiveness with respect to problems reported on the standardized measures. Figure 1 illustrates this point with distributions of CBCL Total Problems scores for

55,508 children from populations that had relatively low, medium, or high mean scores. Even when contrasts between population samples were emphasized by aggregating them into these three groups, large percentages of children in each of the three groups obtained scores that were similar to the scores obtained by large percentages of children in the other two groups.

Research on immigrants

In addition to findings within and between various populations, dimensional assessment research has also yielded important findings for immigrant groups. The most programmatic research of this kind has compared problems reported for Turkish immigrant children in the Netherlands, Turkish children in Turkey, Moroccan immigrant children in the Netherlands, and indigenous Dutch children in the Netherlands (Bengi-Arslan, Verhulst, van der Ende, & Erol, 1997; Crijnen, Bengi-Arslan, & Verhulst, 2000; Murad, Joung, van Lenthe, Bengi-Arslan, & Crijnen, 2003; Stevens et al., 2003). This research provides a model for studies of problems that may be specifically associated with immigration versus problems associated with cultural group and with interactions between effects of immigration and cultural group on problems reported by parents, teachers, and children.

Practical applications of multicultural dimensionally based assessment

The ultimate goal of research on child psychopathology is, of course, to help children. Dimensional instruments that are readily used in diverse contexts, are easily understood by practitioners, and are normed for diverse cultural groups can greatly extend the reach of those who work with troubled children. Computers enable practitioners to display assessment results in relation to normative data derived from research in many populations. Consequently, practitioners can evaluate problems reported by parents, teachers, and children in relation to age, gender, and informant-specific norms for various cultural groups.

The multicultural research findings enable practitioners to tailor their choices of age, gender, and informant-specific norms to the cultural backgrounds of parents, teachers, and children who complete assessment forms. For example, a practitioner working with immigrant children can elect to have scales scored from parent, teacher, and self-ratings displayed in relation to norms appropriate for the child's home society and for the host society. Practitioners working with children who are not immigrants can also benefit from being able to choose norms appropriate for the indigenous culture, rather than being limited to norms from a culture of unknown relevance. Another benefit of easily

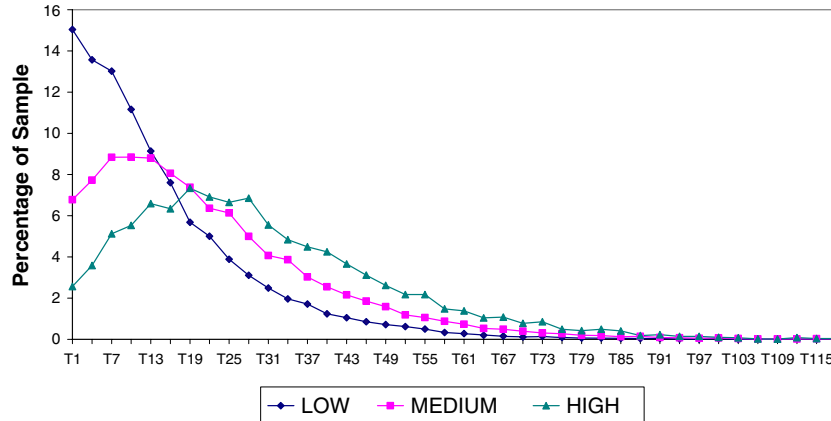


Figure 1 Total Problems score distributions for Low, Medium, and High scoring samples. Each mark on the horizontal axis includes three Total Problems scores; e.g., T1 = 0, 1, 2. *N* = 55,508. (From Achenbach & Rescorla, 2007a).

accessible norms derived from multicultural research is that practitioners working in settings such as refugee camps can apply norms deemed appropriate for children whose cultural backgrounds differ from the practitioners’ own cultural backgrounds.

The easy accessibility of computerized multicultural norms can improve training by helping trainees understand the quantitative variations in problems reported by different informants for children of different backgrounds. Equally important, trainees can quickly see the relativistic nature of clinical cut-points when they view ratings of the same child by different informants in relation to norms for different cultural groups. For example, a child’s scores that are clinically deviant in relation to one set of norms may not be clinically deviant in relation to a different set of norms.

Future directions

The research completed to date establishes a firm foundation on which to base new multicultural research, as outlined in the following sections.

Advancing knowledge

Multicultural dimensionally based research can advance our knowledge in a variety of ways. One way is by obtaining normative data on additional cultural groups (e.g. Woerner et al. 2007) in order to (a) determine whether exceptions to the existing results will be found and (b) expand the pool of normative data available to researchers, practitioners, and trainees. Rigorous multicultural comparisons of SDQ data are especially needed to advance beyond citations of various cutpoint scores from different samples in different studies.

A second way to advance knowledge is by comparing findings for various groups of immigrant children with findings for compatriots who remain in their home cultures and for children indigenous to the

host societies where the immigrant children reside. As mentioned previously, programmatic studies in the Netherlands provide excellent models for research on immigrant children (e.g., Bengi-Arslan et al., 1997; Stevens et al., 2003), as well as longitudinal comparisons of immigrants and nonimmigrants into adulthood (Van Oort et al., 2007). Because millions of immigrant and refugee children reside in host societies that are ill-equipped to evaluate and help them, this is an especially urgent line of research. The need to compare findings for immigrant and refugee children with norms both for their home societies and for their host societies also adds urgency to obtaining normative data for more populations.

A third way to advance knowledge is by undertaking emic studies of the reasons why etic findings that are consistent across most populations are not replicated in one or a few populations. As an example, parents and teachers rate boys higher than girls on statistically derived and DSM-oriented ADHD scales in most populations (Rescorla et al., 2007a, c). An exception is Iran, where the mean scores have been nearly identical for both genders in large, representative samples. Unlike children in the other populations, the Iranian children all attended single-gender schools. Iranian girls obtained higher ADHD scores than girls in most other populations, whereas Iranian boys’ scores were at about the middle of the overall multicultural distribution. We could therefore hypothesize that the absence of boys from the Iranian girls’ classrooms lowered teachers’ thresholds for endorsing ADHD items for girls, thereby producing higher scale scores than found in populations where children of both genders attend the same classrooms.

Unfortunately for this hypothesis, Iranian parents also rated girls as high as boys on statistically derived and DSM-oriented scales of ADHD problems, unlike parents in the other populations. The similar findings for Iranian parent and teacher ratings indicate cross-informant and cross-situational consistency in Iranian adults’ tendency to rate Iranian

girls as high as boys on ADHD problems. Spurred by findings like this, emic studies may reveal differences in etiological factors in different populations. Findings on etiological factors can, in turn, advance our overall understanding of the interplays between biological (including genetic) and environmental (including cultural) influences on phenotypic problems. To round out the picture regarding Iranian ADHD scores, it is interesting to note that Iranian boys and girls self-reported similar levels of ADHD problems (Rescorla et al., 2007b). This lack of gender differences in self-reported ADHD problems was consistent across other societies as well.

Advancing diagnostic assessment

Diagnostic assessment of children's functioning requires face-to-face interviews. The ASEBA has engendered the Semistructured Clinical Interview for Children and Adolescents (SCICA; McConaughy, 2005; McConaughy & Achenbach, 1994, 2001), while the SDQ has engendered the DAWBA (Goodman et al., 2000). The SCICA is administered by a clinical interviewer who asks semistructured questions about various areas of functioning. Based on what the child does and says during the SCICA, the interviewer rates observational and self-report items. The ratings are summed to yield scores on statistically derived syndromes and DSM-oriented scales.

For the DAWBA, lay interviewers administer structured questions that are 'closely related to DSM-IV and ICD-10 diagnostic criteria' (Fleitlich-Bilyk & Goodman, 2004, p. 729). Clinicians then make diagnoses by reviewing printouts of the interviewees' answers. A comparison has been published for the prevalence of DAWBA DSM-IV diagnoses in samples from the UK and the municipality of Taubaté, Brazil, but we know of no direct multicultural comparisons of the prevalence of child diagnoses made from interviews that literally operationalize DSM or ICD criteria.

Several factors may explain the dearth of multicultural research on operationalized diagnoses of children's problems. One factor may be the great cost of having highly trained interviewers administer lengthy interviews to population samples of children and their parents. A second factor may be the need to precisely tailor interview questions to categorical nosological criteria that may be quite remote from the ways in which children and their parents think about the children's problems. A third factor may be that, after questions have been precisely tailored to one set of nosological criteria, the criteria soon change. The changes in DSM criteria for childhood disorders at 7-year intervals from DSM-III in 1980 to DSM-III-R in 1987 and DSM-IV in 1994 (American Psychiatric Association, 1980, 1987, 1994) left insufficient time to develop successive editions of diagnostic interviews, to thoroughly test them, to translate them, and to assemble the massive funding

and research infrastructure needed to uniformly assess representative samples in multiple populations.

An additional impediment may be the present-versus-absent definitions of each nosological criterion and of each diagnosis. The need to first make yes-or-no decisions about each criterion and then about each diagnosis presents practical obstacles to obtaining valid diagnostic data from children and parents who cannot give accurate yes-or-no answers to questions about the children's problems. Requirements for present-versus-absent data also reduce statistical power below what can be obtained with quantified information about problems and may preclude analyses that can simultaneously test quantitative interactions among many variables across samples from many populations, using data from multiple informants. Other adverse consequences of present-versus-absent data include the need for much larger samples than are required for quantitative analyses and the greater difficulty of detecting small effects and interactions.

Dimensionalization of diagnostic criteria

It is possible that the drawbacks of present-versus-absent definitions of criteria and diagnoses will be mitigated in DSM-V, as the American Psychiatric Association has formed a task force to make recommendations for a dimensional approach to DSM-V diagnoses (Helzer, Kraemer, Krueger, Wittchen, & Regier, in press). Even if nosological constructs and criteria continue to be formulated mainly in a top-down fashion, dimensionalization of criteria could have many benefits. The greater differentiation and statistical power afforded by dimensionalization may improve the reliability of standardized diagnostic interviews. Dimensionalization may also improve diagnostic agreement between different interviews and with clinical diagnoses above the poor levels of agreement that have been found in meta-analyses (Rettew et al., 2008). An additional benefit of dimensionalization is that it would facilitate experimentation with different ways of aggregating diagnostic criteria rather than restricting them to fixed rules for making present-versus-absent decisions.

It is important to remember that dimensionalization does not sacrifice the possible benefits of categorization, because categorical cutpoints can be imposed on distributions of dimensionalized scores in order to make present-versus-absent decisions. Consequently, another potential benefit of dimensionalization is that, by assessing large samples of children from multiple cultures with dimensionalized nosological criteria and dimensionally based instruments, researchers can compare the prevalence, patterning, and correlates found with categorical as well as dimensional approaches.

Summary and conclusions

The 21st century presents both opportunities and challenges for helping the world's children. Research, services, and information technology are advancing in beneficial ways. However, customary models for conceptualizing and assessing psychopathology are being challenged by the mixing and clashing of many very different cultures.

We have focused on two sets of instruments that have been used to assess children's functioning in diverse cultural contexts, the ASEBA and the SDQ. Both sets of instruments include parent, teacher, and self-report forms for rating children's problems and favorable characteristics. In contrast to the categorical approach that has been mainly top-down, the ASEBA and SDQ employ a dimensional approach whereby quantitative ratings of items are summed to yield quantitative scale scores. Assessment of individual children can thereby yield profiles of scale scores that can be displayed in relation to norms. The use of the dimensional instruments to assess normative samples in many populations enables users to compare a child's scale scores with age- and gender-specific norms for relevant cultural groups, separately for parent, teacher, and self-ratings. The ease and economy of obtaining quantitative multi-informant data in many languages can greatly facilitate evaluations of children from diverse cultural backgrounds. Equally important, the cumulative findings, mainly from research with the ASEBA in many cultures but augmented by a growing body of SDQ research, provide a firm foundation on which to build future research, services, and training.

In addition to findings specific to particular cultural groups, the findings support the following broad conclusions:

1. Parents, teachers, and youth from very diverse cultural groups are willing and able to complete the dimensional assessment forms, either independently or in interviews.
2. Mean scores on problem scales span a relatively narrow range across very diverse societies, with most mean scores clustered in the middle of the range and only a few societies having relatively low or high mean scores.
3. The distributions of problem scores from all societies overlap considerably with the distributions from all other societies. Moreover, the differences in scores within each society are greater than the differences between the mean scores of different societies.
4. Similar age, gender, and SES effects on problem scores are found in many populations.
5. Uniform CFA procedures support a common set of statistically derived syndromes in many societies.
6. Applications of dimensional assessment make it possible to identify problems that differ for

immigrant children, for their compatriots who reside in the home society, for other immigrant groups, and for children indigenous to the host society.

7. Population samples from many societies provide a basis for multicultural norms that researchers, practitioners, and trainees can readily access by computer.

The multicultural findings set the stage for new efforts such as the following:

1. To obtain normative data on additional cultural groups in order to determine whether exceptions to the existing results will be found and to expand the pool of normative data available to researchers, practitioners, and trainees.
2. To compare findings for more immigrant groups residing in more host societies.
3. To undertake emic studies of reasons for exceptions to etic findings that are consistent across most societies.
4. To advance diagnostic assessment by dimensionalizing diagnostic criteria and assessment procedures.

In sum, a great deal has already been learned from multicultural research employing dimensional assessment. A particularly valuable product of this work is the foundation it provides for many more advances in multicultural research, practice, and training.

Acknowledgement

Many thanks to Prof. Robert Goodman (London) who constructively commented on earlier drafts of the manuscript.

Correspondence to

A. Rothenberger, Child and Adolescent Psychiatry, University of Goettingen, Von-Siebold-Str. 5, D – 37075 Goettingen, Germany; Tel: +49(0)551-396727; Fax: +49(0)551-8120; Email: arothen@gwdg.de

References

- Achenbach, T.M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs*, 80 (No. 615).
- Achenbach, T.M. (1978). The Child Behavior Profile: I. Boys aged 6–11. *Journal of Consulting and Clinical Psychology*, 46, 478–488.
- Achenbach, T.M. (1991a). *Manual for the Child Behavior Checklist/4–18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T.M. (1991b). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

- Achenbach, T.M. (1991c). *Manual for the Youth Self-Report and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T.M., Bird, H.R., Canino, G.J., Phares, V., Gould, M., & Rubio-Stipec, M. (1990). Epidemiological comparisons of Puerto Rican and U.S. mainland children: Parent, teacher, and self reports. *Journal of the American Academy of Child and Adolescent Psychiatry*, 29, 84–93.
- Achenbach, T.M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T.M., & Edelbrock, C. (1986). *Manual for the Teacher's Report Form and Teacher Version of the Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T.M., & Edelbrock, C. (1987). *Manual for the Youth Self-Report and Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T.M., Krukowski, R.A., Dumenci, L., & Ivanova, M.Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, 131, 361–382.
- Achenbach, T.M., & Lewis, M. (1971). A proposed model for clinical research and its application to encopresis and enuresis. *Journal of the American Academy of Child Psychiatry*, 10, 535–554.
- Achenbach, T.M., McConaughy, S.H., & Howell, C.T. (1987). Child/adolescent behavioural and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101, 213–232.
- Achenbach, T.M., & Rescorla, L.A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.
- Achenbach, T.M., & Rescorla, L.A. (2007a). *Multicultural supplement to the Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont Research Center for Children, Youth, and Families.
- Achenbach, T.M., & Rescorla, L.A. (2007b). *Multicultural understanding of child and adolescent psychopathology: Implications for mental health assessment*. New York: Guilford Press.
- Achenbach, T.M., Verhulst, F.C., Edelbrock, C., Baron, G.D., & Akkerhuis, G.W. (1987). Epidemiological comparisons of American and Dutch children: II. Behavioural/emotional problems reported by teachers for ages 6 to 11. *Journal of the American Academy of Child and Adolescent Psychiatry*, 26, 326–332.
- Ambrosini, P.J. (2000). Historical development and present status of the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS). *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 49–58.
- American Psychiatric Association. (1952, 1968, 1980, 1987, 1994). *Diagnostic and statistical manual of mental disorders* (1st edn, 2nd edn, 3rd edn, 3rd edn rev., 4th edn). Washington, DC, Author.
- Arsenault, L., Moffitt, T.E., Caspi, A., Taylor, A., Rijdsdijk, F.V., Jaffee, S.R., et al. (2003). Strong genetic effects on cross-situational antisocial behaviour among 5-year-old children according to mothers, teachers, examiner-observers, and twins' self-reports. *Journal of Child Psychology and Psychiatry*, 44, 832–848.
- Auerbach, J.G., & Lerner, Y. (1991). Syndromes derived from the Child Behavior Checklist for clinically referred Israeli boys aged 6–11. *Journal of Child Psychology and Psychiatry*, 32, 1017–1024.
- Becker, A., Hagenberg, N., Roessner, V., Woerner, W., & Rothenberger, A. (2004). Evaluation of the self-reported SDQ in a clinical setting: Do self-reports tell us more than ratings by adult informants? *European Child and Adolescent Psychiatry*, 13(Suppl. 2), 17–24.
- Becker, A., Steinhausen, H.C., Balursson, G., Dalsgaard, S., Lorenzo, M.J., Ralston, S.J., et al. (2006). Psychopathological screening of children with ADHD: Strengths and Difficulties Questionnaire in a pan-European study. *European Child and Adolescent Psychiatry*, 15(Suppl. 1), 56–62.
- Becker, A., Woerner, W., Hasselhorn, M., Banaschewski, T., & Rothenberger, A. (2004). Validation of the parent and teacher SDQ in a clinical sample. *European Child and Adolescent Psychiatry*, 13(Suppl. 2), 11–16.
- Bengi-Arslan, L., Verhulst, F.C., van der Ende, J., & Erol, N. (1997). Understanding childhood (problems) behaviors from a cultural perspective: Comparison of problem behaviors and competencies in Turkish immigrant, Turkish and Dutch children. *Social Psychiatry and Psychiatric Epidemiology*, 32, 477–484.
- Bilenberg, N. (1999). The Child Behavior Checklist (CBCL) and related material: Standardization and validation in Danish population and clinically based samples. *Acta Psychiatrica Scandinavica, Supplementum*, 100, 398, 1–52.
- Bird, H. (1996). Epidemiology of childhood disorders in cross-cultural context. *Journal of Child Psychology and Psychiatry*, 37, 35–49.
- Birmaher, B., Stanley, M., Greenhill, L., Twomey, J., Gavrilescu, A., & Rabinovich, H. (1990). Platelet imipramine binding in children and adolescents with impulsive behavior. *Journal of the American Academy of Child and Adolescent Psychiatry*, 29, 914–918.
- Bourdon, K.H., Goodman, R., Rae, D.S., Simpson, G., & Koretz, D.S. (2005). The Strengths and Difficulties Questionnaire: U.S. normative data and psychometric properties. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44, 557–564.
- Browne, N.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edn). New York: Academic Press.
- Conners, C.K. (1997). *Conners' Rating Scales-Revised technical manual*. North Tonawanda, NY: Multi-Health Systems.
- Crijnen, A.A.M., Achenbach, T.M., & Verhulst, F.C. (1997). Comparisons of problems reported by parents of children in 12 cultures: Total Problems, Externalizing, and Internalizing. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36, 1269–1277.

- Crijnen, A.A.M., Achenbach, T.M., & Verhulst, F.C. (1999). Comparisons of problems reported by parents of children in twelve cultures: The CBCL/4–18 syndrome constructs. *American Journal of Psychiatry*, *156*, 569–574.
- Crijnen, A.A.M., Bengi-Arslan, L., & Verhulst, F.C. (2000). Teacher-reported problem behaviour in Turkish immigrant and Dutch children: A cross-cultural comparison. *Acta Psychiatrica Scandinavica*, *102*, 439–444.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- De Groot, A., Koot, H.M., & Verhulst, F.C. (1994). Cross-cultural generalizability of the CBCL cross-informant syndromes. *Psychological Assessment*, *6*, 225–230.
- De Groot, A., Koot, H.M., & Verhulst, F.C. (1996). Cross-cultural generalizability of the Youth Self-Report and Teacher's Report Form cross-informant syndromes. *Journal of Abnormal Child Psychology*, *24*, 651–664.
- De Los Reyes, A., & Kazdin, A.E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483–509.
- Dickey, W.C., & Blumberg, S.J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 1159–1167.
- Döpfner, M., Berner, W. & Lehmkuhl, G. (1995a). Reliabilität und faktorielle Validität der Youth Self-Report der Child Behavior Checklist bei einer klinischen Stichprobe. *Diagnostica*, *41*, 221–244.
- Döpfner, M., Berner, W. & Lehmkuhl, G. (1997). Verhaltensauffälligkeiten von Schülern im Urteil der Lehrer – Reliabilität und faktorielle Validität der Teacher's Report Form der Child Behavior Checklist. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *18*, 199–214.
- Döpfner, M., Berner, W., Schmeck, K., Lehmkuhl, G., & Poustka, F. (1995b). Internal consistency and validity of the CBCL and the TRF in a German sample – a cross cultural comparison. In J. Sergeant (Ed.), *Eunethydis. European approaches to hyperkinetic disorder* (pp. 51–81). Zürich: Fotorotar.
- Döpfner, M., Schmeck, K., Berner, W., Lehmkuhl, G. & Poustka, F. (1994). Zur Reliabilität und faktoriellen Validität der Child Behavior Checklist – eine Analyse in einer klinischen und einer Feldstichprobe. *Zeitschrift für Kinder- und Jugendpsychiatrie*, *22*, 189–205.
- Edelbrock, C., Rende, R., Plomin, R., & Thompson, L.A. (1995). A twin study of competence and problem behavior in childhood and early adolescence. *Journal of Child Psychology and Psychiatry*, *36*, 775–785.
- Eley, T.C., Lichtenstein, P., & Moffitt, T.E. (2003). A longitudinal behavioural genetic analysis of the etiology of aggressive and nonaggressive antisocial behavior. *Development and Psychopathology*, *15*, 383–402.
- Eley, T.C., Lichtenstein, P., & Stevenson, J. (1999). Sex differences in the etiology of aggressive and nonaggressive antisocial behavior: Results from two twin studies. *Child Development*, *70*, 155–168.
- Fleitlich-Bilyk, B., & Goodman, R. (2004). The prevalence of child psychiatric disorders in Southeast Brazil. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 727–734.
- Fombonne, E. (1992). Parent reports on behaviour and competencies among 6–11-year-old French children. *European Child and Adolescent Psychiatry*, *1*, 233–243.
- Gabel, S., Stadler, J., Bjorn, J., Shindledacker, R., & Bowden, C. (1993). Dopamine-beta-hydroxylase in behaviourally disturbed youth. Relationship between teacher and parent ratings. *Biological Psychiatry*, *34*, 434–442.
- Ghodsian-Carpey, J., & Baker, L.A. (1987). Genetic and environmental influences on aggression in 4- to 7-year-old twins. *Aggressive Behavior*, *13*, 173–186.
- Gjone, H., Stevenson, J., & Sundet, J.M. (1996). Genetic influence on parent-reported attention-related problems in a Norwegian general population twin sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 588–596.
- Goodman, R. (1994). A modified version of the Rutter Parent Questionnaire including extra items on children's strengths: A research note. *Journal of Child Psychology and Psychiatry*, *35*, 1483–1494.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581–586.
- Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology and Psychiatry*, *40*, 791–799.
- Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 1337–1345.
- Goodman, R., Ford, T., Corbin, T., & Meltzer, H. (2004). Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. *European Child and Adolescent Psychiatry*, *13*, 25–31.
- Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *41*, 645–655.
- Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, *177*, 534–539.
- Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties scale: A pilot study on the self-report version. *European Child and Adolescent Psychiatry*, *7*, 125–130.
- Goodman, R., Renfrew, D., & Mullick, M. (2000). Predicting type of psychiatric disorder from Strengths and Difficulties Questionnaire (SDQ) scores in child mental health clinics in London and Dhaka. *European Child and Adolescent Psychiatry*, *9*, 129–134.
- Goodman R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, *27*, 17–24.

- Gove, P. (Ed.). (1971). *Webster's third new international dictionary of the English language*. Springfield, MA: Merriam.
- Gregory, A.M., Eley, T.C., & Plomin, R. (2004). Exploring the association between anxiety and conduct problems in a large sample of twins aged 2–4. *Journal of Abnormal Child Psychology*, 33, 111–122.
- Hanna, L. (1995). Demographic and clinical features of obsessive-compulsive disorder in children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 34, 19–27.
- Hartman, C.A., Hox, J., Auerbach, J., Erol, N., Fonseca, A.C., Mellenbergh, G.J., et al. (1999). Syndrome dimensions of the Child Behavior Checklist and the Teacher Report Form: A critical empirical evaluation. *Journal of Child Psychology and Psychiatry*, 40, 1095–1116.
- Heijmans Visser, J., van der Ende, J., Koot, H.M., & Verhulst, F.C. (2000). Predictors of psychopathology in young adults referred to mental health services in childhood or adolescence. *British Journal of Psychiatry*, 177, 59–65.
- Helstälä, L., Sourander, A., & Bergroth, L. (2001). Parent-reported competence and emotional and behavioural problems in Finnish adolescents. *Nordic Journal of Psychiatry*, 55, 337–341.
- Helzer, J., Kraemer, H., Krueger, R.F., Wittchen, H.U., & Regier, D.A. (Eds.). (in press). *Dimensional approaches in diagnostic classification: Refining the research agenda for DSM-V*. Washington, DC: American Psychiatric Association.
- Hermans, H.J.M., & Kempen, H.J.G. (1998). Moving cultures: The perilous problems of cultural dichotomies in a globalizing society. *American Psychologist*, 53, 1111–1120.
- Heubeck, B.G. (2000). Cross-cultural generalizability of CBCL syndromes across three continents: From the USA and Holland to Australia. *Journal of Abnormal Child Psychology*, 28, 439–450.
- Hofstra, M.B., van der Ende, J., & Verhulst, F.C. (2002). Child and adolescent problems predict DSM-IV disorders in adulthood: A 14-year follow-up of a Dutch epidemiological sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 41, 182–189.
- Hysing, M., Elgen, I., Gillberg, C., Lie, S.A., & Lunder-vold, A.J. (2007). Chronic physical illness and mental health in children. Results from a large-scale population study. *Journal of Child Psychology and Psychiatry*, 48, 785–792.
- Ickowicz, A., Schachar, R.J., Sugarman, R., Chen, S.X., Millette, C., & Cook, L. (2006). The parent interview for child symptoms: A situation-specific clinical research interview for attention-deficit hyperactivity and related disorders. *Canadian Journal of Psychiatry*, 51, 325–328.
- Ivanova, M.Y., Achenbach, T.M., Dumenci, L., Rescorla, L.A., Almqvist, F., Bilenberg, N., et al. (2007a). Testing the 8-syndrome structure of the CBCL in 30 societies. *Journal of Clinical Child and Adolescent Psychology*, 36, 405–417.
- Ivanova, M.Y., Achenbach, T.M., Rescorla, L.A., Dumenci, L., Almqvist, F., Bathiche, M., et al. (2007b). Testing the Teacher's Report Form syndromes in 20 societies. *School Psychology Review*, 36, 468–483.
- Ivanova, M.Y., Achenbach, T.M., Rescorla, L.A., Dumenci, L., Almqvist, F., Bilenberg, N., et al. (2007c). The generalizability of the Youth Self-Report syndrome structure in 23 societies. *Journal of Consulting and Clinical Psychology*, 75, 729–738.
- Kashala, E., Elgen, I., Sommerfelt, K., & Tylleskar, T. (2005). Teacher ratings of mental health among school children in Kinshasa, Democratic Republic of Congo. *European Child and Adolescent Psychiatry*, 12, 208–215.
- Klasen, H., Woerner, W., Wolke, D., Meyer, R., Overmeyer, S., Kaschnitz, W., et al. (2000). Comparing the German versions of the Strengths and Difficulties Questionnaire (SDQ-Deu) and the Child Behavior Checklist. *European Child and Adolescent Psychiatry*, 9, 271–276.
- Koskelainen, M., Sourander, A., & Kaljonen, A. (2000). The Strengths and Difficulties Questionnaire among Finnish school-aged children and adolescents. *European Child and Adolescent Psychiatry*, 9, 277–284.
- Koskelainen, M., Sourander, A., & Vauras, M. (2001). Self-reported strengths and difficulties in a community sample of Finnish adolescents. *European Child and Adolescent Psychiatry*, 10, 180–185.
- Leung, P.W.L., Kwong, S.L., Tang, C.P., Ho, T.P., Hung, S.F., Lee, C.C., et al. (2006). Test-retest reliability and criterion validity of the Chinese version of CBCL, TRF, and YSR. *Journal of Child Psychology and Psychiatry*, 47, 970–973.
- Liu, X., Kurita, H., Guo, C., Tachimori, H., Ze, J., & Okawa, M. (2000). Behavioural and emotional problems in Chinese children: Teacher reports for ages 6 to 11. *Journal of Child Psychology and Psychiatry*, 41, 253–260.
- MacDonald, V.M., & Achenbach, T.M. (1999). Attention problems versus conduct problems as six-year predictors of signs of disturbance in a national sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 38, 1254–1261.
- McConaughy, S.H. (2005). *Clinical interviews for children and adolescents: Assessment to intervention*. New York: Guilford Press.
- McConaughy, S.H., & Achenbach, T.M. (1994, 2001). *Manual for the Semistructured Clinical Interview for Children and Adolescents* (1st & 2nd edn). Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.
- Martin, N., Scourfield, J., & McGuffin, P. (2002). Observer effects and the heritability of childhood attention-deficit hyperactivity disorder symptoms. *British Journal of Psychiatry*, 180, 260–265.
- Mellor, D. (2004). Furthering the use of the Strengths and Difficulties Questionnaire: Reliability with younger child respondents. *Psychological Assessment*, 16, 396–401.
- Montenegro, H. (1983). *Salud mental del escolar. Estandarización del inventario de problemas conductuales y destrezas sociales de T. Achenbach en niños de 6 a 11 años*. Santiago, Chile: Centro de Estudios de Desarrollo y Estimulación Psicosocial.
- Murad, S.D., Joung, I.M.A., van Lenthe, F.J., Bengi-Arslan, L., & Crijnen, A.A.M. (2003). Predictors of self-reported problem behaviours in Turkish immigrant and Dutch adolescents in the Netherlands.

- Journal of Child Psychology and Psychiatry*, 44, 412–423.
- Muris, P., Meesters, C., Eijkelenboom, A., & Vincken, M. (2004). The self-report version of the Strengths and Difficulties Questionnaire: Its psychometric properties in 8- to 13-year-old non-clinical children. *British Journal of Clinical Psychology*, 43, 437–448.
- Obel, C., Heiervang, E., Rodriguez, A., Heyerdahl, S., Smedje, H., Sourander, A., et al. (2004). The Strengths and Difficulties Questionnaire in the Nordic countries. *European Child and Adolescent Psychiatry*, 13(Suppl. 2), 32–39.
- Pike, K.L. (1954). *Language in relation to a unified theory of the structure of human behavior*. Glendale, CA: Summer Institute of Linguistics.
- Pike, K.L. (1967). *Language in relation to a unified theory of the structure of human behavior*. The Hague: Mouton.
- Rescorla, L.A., Achenbach, T.M., Ginzburg, S., Ivanova, M.Y., Dumenci, L., Almqvist, F., et al. (2007a). Consistency of teacher-reported problems for students in 21 countries. *School Psychology Review*, 36, 91–110.
- Rescorla, L.A., Achenbach, T.M., Ivanova, M.Y., Dumenci, L., Almqvist, F., Bilenberg, N., et al. (2007b). Epidemiological comparisons of problems and positive qualities reported by adolescents in 24 countries. *Journal of Consulting and Clinical Psychology*, 75, 351–358.
- Rescorla, L.A., Achenbach, T.M., Ivanova, M.Y., Dumenci, L., Almqvist, F., Bilenberg, N., et al. (2007c). Behavioural and emotional problems reported by parents of children ages 6 to 16 in 31 societies. *Journal of Emotional and Behavioural Disorders*, 15, 130–142.
- Rettew, D.C., Doyle, A.C., Achenbach, T.M., Dumenci, L., & Ivanova, M.Y. (2008). Meta-analyses of diagnostic agreement between clinical evaluations and standardized diagnostic interviews. In review.
- Reynolds, C.R., & Kamphaus, R.W. (1992). *Behavior Assessment System for Children Parent Rating Scales*. Circle Pines, MN: American Guidance Service.
- Rietveld, M.J.H., Hudziak, J.J., Bartels, M., Van Beijsterveldt, C.E.M., & Boomsma, D.I. (2004). Heritability of attention problems in children: Longitudinal results from a study of twins, age 3 to 12. *Journal of Child Psychology and Psychiatry*, 45, 577–588.
- Rønning, J.A., Handegaard, B.H., Sourander, A., & Mørch, W-T. (2004). The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. *European Child and Adolescent Psychiatry*, 13, 73–82.
- Roessner, V., Becker, A., Rothenberger, A., Rohde, L.A., Banaschewski, T. (2007). A cross-cultural comparison between samples of Brazilian and German children with ADHD/HD using Child Behavior Checklist. *European Archives of Psychiatry and Critical Neuroscience* 257, 352–359.
- Rothenberger, A., & Woerner, W. (Eds.). (2004). Strength and Difficulties Questionnaire (SDQ) – Evaluations and Applications. *European Child and Adolescent Psychiatry*, 13(Suppl. 2), ii1–ii2.
- Ruchkin, V., Koposov, R., & Schwab-Stone, M. (2007). The Strength and Difficulties Questionnaire: Scale validation with Russian adolescents. *Journal of Clinical Psychology*, 63, 861–869.
- Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: Preliminary findings. *Journal of Child Psychology and Psychiatry*, 8, 1–11.
- Rutter, M., Tizard, J., & Whitmore, K. (1970). *Education, health and behaviour*. New York: Wiley.
- Sawyer, M.G., Mudge, J., Carty, V., Baghurst, P., & McMichael, A. (1996). A prospective study of childhood emotional and behavioural problems in Port Pirie, South Australia. *Australian and New Zealand Journal of Psychiatry*, 30, 781–787.
- Schmeck, K., Poustka, F., Döpfner, M., Plücker, J., Berner, W., Lehmkuhl, G., et al. (2001). Discriminant validity of the Child Behaviour Checklist CBCL-4/18 in German samples. *European Child and Adolescent Psychiatry*, 10, 240–247.
- Scourfield, J., Van den Bree, M., Martin, N., & McGuffin, P. (2004). Conduct problems in children and adolescents: A twin study. *Archives of General Psychiatry*, 61, 489–496.
- Shaffer, D., Fisher, P., Lucas, C.P., Dulcan, M.K., & Schwab-Stone, M.E. (2000). NIMH Diagnostic Interview Schedule for Children version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 28–38.
- Sims, A., Mundt, C., Berner, P., Barocka, A. (2000). Descriptive phenomenology. In M.G. Gelder, J.P. Lopez-Ibor Jr, N.C. Andreasen (Eds.), *New Oxford Textbook of Psychiatry* (pp. 56–79). Oxford: Oxford University Press.
- Smedje, H., Broman, J.-E., Hetta, J., & von Knorring, A.-L. (1999). Psychometric properties of a Swedish version of the 'Strengths and Difficulties Questionnaire'. *European Child and Adolescent Psychiatry*, 8, 63–70.
- Stanger, C., Achenbach, T.M., & Verhulst, F.C. (1997). Accelerated longitudinal comparison of aggressive versus delinquent syndromes. *Development and Psychopathology*, 9, 43–58.
- Stanger, C., MacDonald, V., McConaughy, S.H., & Achenbach, T.M. (1996). Predictors of cross-informant syndromes among children and youths referred for mental health services. *Journal of Abnormal Child Psychology*, 24, 597–614.
- Stevens, G.W.J.M., Pels, T., Bengi-Arslan, L., Verhulst, F.C., Vollebergh, W.A.M., & Crijnen, A.A.M. (2003). Parent, teacher and self-reported problem behavior in The Netherlands: Comparing Moroccan immigrant with Dutch and with Turkish immigrant children and adolescents. *Social Psychiatry and Psychiatric Epidemiology*, 38, 576–585.
- Stoff, D.M., Pollock, L., Vitiello, B., Behar, D., & Bridger, W.H. (1987). Reduction of 3-H-imipramine binding sites on platelets of conduct disordered children. *Neuropsychopharmacology*, 1, 55–62.
- Taylor, E., Schachar, R., Thorley, G., & Wieselberg, M. (1986). Conduct disorder and hyperactivity: Separation of hyperactivity and antisocial conduct in British child psychiatric patients. *British Journal of Psychiatry*, 149, 760–770.
- Triandis, H.C. (1989). The self and social behavior in differing cultural contexts. *Psychological Review*, 96, 506–520.
- van den Oord, E.J.C.G., Verhulst, F.C., & Boomsma, D.I. (1996). A genetic study of maternal and

- paternal ratings of problem behaviors in three-year-old twins. *Journal of Abnormal Psychology*, 105, 349–357.
- van der Valk, J.C., van den Oord, E.J.C.G., Verhulst, F.C., & Boomsma, D.I. (2001). Using parental ratings to study the etiology of 3-year-old twins' problem behaviors: Different views or rater bias? *Journal of Child Psychology and Psychiatry*, 42, 921–931.
- Van Oort, F.V.A., Joung, I.M.A., Mackenbach, J.P., Verhulst, F.C., Bengi-Arslan, L., Crijnen, A.A.M., et al. (2007). Development of ethnic disparities in internalizing and externalizing problems from adolescence into young adulthood. *Journal of Child Psychology and Psychiatry*, 48, 176–184.
- Van Widenfelt, B.M., Goedhart, A.W., Treffers, P.D.A., & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European Child and Adolescent Psychiatry*, 12, 281–289.
- Verhulst, F.C., Achenbach, T.M., van der Ende, J., Erol, N., Lambert, M.C., Leung, P.W.L., et al. (2003). Comparisons of problems reported by youths from seven countries. *American Journal of Psychiatry*, 160, 1479–1485.
- Verhulst, F.C., Akkerhuis, G.W., & Althaus, M. (1985). Mental health in Dutch children: (I) A cross-cultural comparison. *Acta Psychiatrica Scandinavica, Supplementum*, 72, 323, 1–108.
- Verhulst, F.C., Prince, J., Vervuurt-Poot, C., & de Jong, J.B. (1989). Mental health in Dutch children: (IV) Self-reported problems for ages 11–18. *Acta Psychiatrica Scandinavica, Supplementum*, 80, 356, 1–48.
- Woerner, W., Becker, A., & Rothenberger, A. (2004). Normative data and scale properties of the German parent SDQ. *European Child and Adolescent Psychiatry*, 13(Suppl. 2), 3–10.
- Woerner, W., Nuanmanee, S., Wongpiromsarn, Y., Goodman, R., Becker, A., Rothenberger, A. (2007). Thai parent-rated Strengths and Difficulties Questionnaire (SDQ): normative data; scale properties, and comparison with European field samples. *Poster, 2nd International Conference on Child and Adolescent Psychopathology*, London 5-6 July.
- World Health Organization. (1992). *Mental disorders: Glossary and guide to their classification in accordance with the Tenth Revision of the International Classification of Diseases* (10th edn). Geneva: World Health Organization.
- Yu, C.Y., & Muthén, B.O. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes (Technical Report)*. Los Angeles: University of California at Los Angeles, Graduate School of Education and Information Studies.

Manuscript accepted 16 October 2007