Diagnostic accuracy of self and parent rating scales in the prediction of psychiatric diagnoses in children and adolescents

Aebi, Marcel

Abstract: Although parent and self-rating scales are often used in clinical and research settings, their accuracy in the prediction of psychiatric disorders according to ICD-10 or DSM-IV often remains unclear. In the present thesis the diagnostic accuracy of three parental rating scales in the prediction of Attention-Deficit-Hyperactivity Disorder (ADHD) and Oppositional Defiant Disorder (ODD) and two self-rating scales in the prediction of adolescent depression were tested in three separate studies. The first study found the recently introduced DSM-oriented attention problem scale of the Child Behavior Checklist (CBCL) more adequate than the previous empirical defined attention problem scale in the identification of ADHD. This was in particular true for subjects in a clinical sample referred for various psychiatric disorders. A cut-off score of 5 was recommended for clinical practice. In a second study a sample of adolescents with clinical depression was compared to a sample of unpreferred community controls. The Youth Self Report (YSR) and the Center of Epidemiological Studies-Depression Scale (CES-D) showed excellent ability in the discrimination of these two samples. A range of acceptable cut-off scores between 5 and 9 on the YSR affective problem scale and between 12 and 31 on the CES-D scale served best in the prediction of clinical depressive episodes in adolescents. In a third study the Conners' Parent Ratings Scale-Revised (CPRS-R) and the parent version of the Strength and Difficulties Questionnaire (PSDQ) were tested in the prediction of ODD in a large transnational sample of ADHD referred children and adolescents. Furthermore, the construct validity of three previously described dimensions of ODD was examined and finally the accuracy of the CPRS-R and the PSDQ were tested in the prediction of these separate ODD dimensions. The CPRS-R oppositional scale and the PSDQ conduct problem scale showed adequate diagnostic accuracy. Furthermore, the construct validity of three ODD dimensions labeled ODD-irritable, ODD-headstrong and ODD-hurtful was confirmed. Furthermore, our results convincingly show that a three factor structure of ODD is more appropriate than a single general factor of ODD. The CPRS-R emotional lability scale was able to predict ODD-irritable significantly. Overall, these three studies confirmed the diagnostic accuracy of clinical rating scales in the prediction of psychiatric disorders in youth. Furthermore, these results are of clinical importance as newer and diagnosis-oriented rating scales showed better diagnostic accuracy and can be recommended for the initial psychiatric assessment of children and adolescents. However, despite the good validity of rating scales, further information on age of onset, continuity, impairment, specificity of symptoms and information about other psychiatric disorders should be included in order to arrive at the final diagnosis. Obschon Selbst- und Fremdbeurteilungsskalen häufig in klinischen Institutionen und in der Forschung eingesetzt werden, ist ihre diagnostische Validität für die Vorhersage psychischer Störungen gemäss ICD-10 oder DSM-IV häufig unklar. In der vorliegenden Doktorarbeit wird die diagnostische Validität von drei Elternbeurteilungsskalen zur Prädiktion von Aufmerksamkeitsdefizit-Hyperaktivitäts-Störungen (ADHS) und von Störungen mit oppositionellem Trotzverhalten (SOT) sowie zwei Selbstbeurteilungsskalen zur Prädiktion von klinischen Depressionen in insgesamt drei verschiedenen Untersuchungen überprüft. Die erste Studie zeigte, dass die kürzlich eingeführte DSM-orientierte Aufmerksamkeitsstörungsskala der „Child Behavior Checklist" (CBCL) genauer ist in der Prädiktion von ADHS als die vorhergehende empirisch definierte Skala Aufmerksamkeitsprobleme. Dies bewahrheitete sich insbesondere in einer klinischen Stichprobe, welche Probanden mit verschiedenen psychischen Störungen beinhaltete. Ein Grenzwert von 5 wurde für die klinische Praxis empfohlen. In einer zweiten

Studie wurde eine Stichprobe von Jugendlichen mit klinischer Depression mit einer Kontrollstichprobe verglichen. Der „Youth Self Report“ (YSR) und die Allgemeine Depressions-Skala (ADS) zeigten eine exzellente Fähigkeit die beiden Stichproben zu diskriminieren. Für die Prädiktion von klinischen Depressionen zeigte sich ein Bereich akzeptabler Grenzwerte zwischen 5 und 9 der DSM-orientierten Skala affektive Schwierigkeiten des YSR, bzw. zwischen 12 und 31 für die ADS, als am besten geeignet. In einer dritten Studie wurde die revidierte Version Conners des Conners-Elternfragebogen (Conners' Parent Ratings Scale Revised; CPRS-R) und die Eltern-Version des Fragebogen zu Stärken und Schwächen (Strength and Difficulties Questionnaire; PSDQ) in Bezug auf ihre prädiktive Validität von SOT in einer internationalen Stichprobe von Kindern und Jugendlichen mit ADHS untersucht. Weiter wurde die Konstruktvalidität von SOT untersucht und schliesslich die prädiktive Validität des CPRS-R und des PSDQ in Bezug auf drei verschiedene Dimensionen der SOT analysiert. Die Skala oppositionelles Verhalten des CPRS-R und die Skala Verhaltensprobleme des PSDQ zeigten eine adäquate diagnostische Validität für SOT. Weiter konnte die Konstruktvalidität der drei verschiedenen Dimensionen der SOT (Irritabilität, Dickköpfigkeit und schädliches Verhalten) bestätigt werden. Die Ergebnisse zeigten überzeugend, dass eine 3-Faktorenstruktur den Daten angemessener ist als ein allgemeiner Faktor von SOT. Die CPRS-R Skala emotionale Labilität war in der Lage den Faktor Irritabilität statistisch signifikant zu bestimmen. Insgesamt bestätigen die vorliegenden drei Untersuchungen die Validität von klinischen Beurteilungsskalen in der Prädiktion von psychischen Störungen im Kindes- und Jugendalter. Die neueren und diagnoseorientierten Skalen zeigten gegenüber anderen Skalen eine überlegene diagnostische Validität und können für klinische Abklärungen von Kindern und Jugendlichen empfohlen werden. Trotz der guten diagnostischen Validität der Selbst- und Fremdbeurteilungsskalen sind weitere Informationen über Beginn, Verlauf, Beeinträchtigung und Spezifität der Symptome sowie über das Vorliegen von weiteren psychischen Störungen notwendig, um zu einer endgültigen Diagnose zu gelangen.

Diagnostic Accuracy of Self and Parent Rating Scales in the Prediction of Psychiatric Diagnoses in Children and Adolescents

Thesis
presented to the Faculty of Arts
of the
University of Zurich
for the degree of Doctor of Philosophy

by

Marcel Aebi
of Heimiswil (Be)

accepted in January 2009 on the recommendation of
Prof. Dr. med. Dr. phil. Hans-Christoph Steinhausen
Prof. Dr. phil. Friedrich Wilkening

Zurich
2009

Abstract

Although parent and self rating scales are often used in clinical and research settings, their accuracy in the prediction of psychiatric disorders according to ICD-10 or DSM-IV often remains unclear. In the present thesis the diagnostic accuracy of three parental rating scales in the prediction of Attention-Deficit-Hyperactivity Disorder (ADHD) and Oppositional Defiant Disorder (ODD) and two self rating scales in the prediction of adolescent depression were tested in three separate studies. The first study found the recently introduced DSM-oriented attention problem scale of the Child Behavior Checklist (CBCL) more adequate than the previous empirical defined attention problem scale in the identification of ADHD. This was in particular true for subjects in a clinical sample referred for various psychiatric disorders. A cut-off score of 5 was recommended for clinical practice. In a second study a sample of adolescents with clinical depression was compared to a sample of unreferred community controls. The Youth Self Report (YSR) and the Center of Epidemiological Studies-Depression Scale (CES-D) showed excellent ability in the discrimination of these two samples. A range of acceptable cut-off scores between 5 and 9 on the YSR affective problem scale and between 12 and 31 on the CES-D scale served best in the prediction of clinical depressive episodes in adolescents. In a third study the Conners' Parent Ratings Scale Revised (CPRS-R) and the parent version of the Strength and Difficulties Questionnaire (PSDQ) were tested in the prediction of ODD in a large transnational sample of ADHD referred children and adolescents. Furthermore, the construct validity of three previously described dimensions of ODD was examined and finally the accuracy of the CPRS-R and the PSDQ were tested in the prediction of these separate ODD dimensions. The CPRS-R oppositional scale and the PSDQ conduct problem scale showed adequate diagnostic accuracy. Furthermore, the construct validity of three ODD dimensions labeled ODD-irritable, ODD-headstrong and ODD-hurtful was confirmed. Furthermore, our results convincingly show that a three factor structure of ODD is more appropriate than a single general factor of ODD. The CPRS-R emotional lability scale was able to predict ODD-irritable significantly. Overall, these three studies confirmed the diagnostic accuracy of clinical rating

scales in the prediction of psychiatric disorders in youth. Furthermore, these results are of clinical importance as newer and diagnosis-oriented rating scales showed better diagnostic accuracy and can be recommended for the initial psychiatric assessment of children and adolescents. However, despite the good validity of rating scales, further information on age of onset, continuity, impairment, specificity of symptoms and information about other psychiatric disorders should be included in order to arrive at the final diagnosis.

Acknowledgement

This doctoral dissertation was realized under the direction of Prof. Dr. Dr. Hans-Christoph Steinhausen. I would like to express my deep appreciation for supporting my work, and for his advice throughout the dissertation process. In particular I am very grateful for his fast and highly profound comments and his skilled editing of my manuscripts. In addition, I would like to thank Prof. Dr. Friedrich Wilkening for supporting this doctoral thesis.

Furthermore, I am also grateful to Robert Goodman, Frank Verhulst, Argyris Stringaris and Steve Faraone for their practical advices, suggestions and the constructive criticisms regarding the three manuscripts which are included in the present thesis.

Data for the present studies were coming from the Zurich Adolescent Psychopathology and Psychology Study (ZAPPS) and the International Multicentre ADHD Genetics (IMAGE) study. I am profoundly indebted to Hans-Christoph Steinhausen, principal investigator, and Christa Winkler, basic coordinator of the ZAPPS study, and to the IMAGE executive committee for making this work possible.

Furthermore I am very grateful to Susanne Eschmann for her statistical advices and to Madleina Manetsch and Gabriela Lazzeri for proofreading parts of the present thesis.

A special tribute goes to the psychiatric patients of the Child and Adolescent Psychiatric Service of the Canton Zurich. For me as clinician, the contacts and discussions with the young people and their parents were a continuous source of inspiration.

Lastly, I am extremely thankful to my parents for enabling me to enter in this fascinating world of research.

Table of Contents

1 Introduction

Evidence-based therapies for children and adolescents are delineated for different emotional and behavioral disorders (e. g. Dubicka & Wilkinson, 2007; Nair, Ehimare, Beitman, Nair, & Lavin, 2006). However, before clinicians can decide about the use of an evidence-based therapy, they must identify problems to target in treatment. Thus, the use of an evidence-based treatment is contingent upon a valid and evidence-based assessment. Emotional and behavioral disorders in children and adolescents are classified in form of psychiatric diagnoses defined by criteria which include multiple symptoms as well as the severity, the onset, duration and impairment of the corresponding disorder. In order to identify psychiatric disorders, diagnostic instruments such as self and parent rating scales were developed. Self and parent rating scales are frequently used in psychiatric assessments because they are easy to use and are time- and cost-saving.

The present thesis deals with the diagnostic accuracy of parent and self rating scales in the prediction of psychiatric diagnoses in children and adolescents. Therefore, the concordance of a test result (e.g. from a rating scale) and an independent measure of psychiatric diagnoses as “gold standard” was tested. For psychiatric diagnoses two separate methods chosen as diagnostic “gold standards” were applied in the studies included in order to arrive at psychiatric diagnoses: (i) standardized diagnostic interviews with a parent and (ii) expert opinions based on various clinical information.

At the beginning of the present thesis, an overview of child and adolescent psychiatric assessments in the light of evidence-based practice is given. Additionally, both of the methods used as “gold standard” in the present thesis are critically discussed. Furthermore, advantages and disadvantages of rating scales in the diagnostic process are shown. The following chapter is dealing with the accuracy of a diagnostic test. Therefore, different methodological approaches for measuring diagnostic accuracy are presented. The main part of the present thesis comprises three different studies concerning diagnostic accuracy of self

and parent rating scales in the prediction of three common psychiatric disorders, e.g. Attention-Deficit-Hyperactivity Disorder (ADHD), clinical depression and Oppositional Defiant Disorder (ODD). In the final discussion, the summarized results of these studies were evaluated. In addition, a general conclusion for the use of rating scales in the psychiatric assessment of children and adolescents is given.

2 Diagnostic assessment in child and adolescent psychiatry

Over the past years, evidence was accumulated to suggest that there is a significant discrepancy between knowledge gained from clinical trials regarding the assessment of mental disorders and the actual psychiatric assessment of children and adolescents in clinical practice (Doss, 2005; A. L. Jensen & Weisz, 2002). Meanwhile, a growing trend for evidence- based practice in child and adolescent mental health can be recognized (American Academy of Child and Adolescent Psychiatry, 2005). Evidence-based guidelines for assessment of various psychiatric disorders were reported (e.g. ADHD, autism, anxiety; Evans & Youngstrom, 2006; Pelham, Fabiano, & Massetti, 2005; Reichow, Volkmar, & Cicchetti, 2008; Silverman & Ollendick, 2005). Furthermore, various standardized instruments for the assessment of mental health problems in children and adolescents are available and are being recommended for clinical practice: First, structured or semi-structured psychiatric interviews with the adolescent or his/her parents are often described as “gold standard” for psychiatric diagnosis. These instruments are mostly based on the criteria of the commonly used classification systems for mental health problems as the ICD-10 (World Health Organization, 1994) or the DSM-IV (American Psychiatric Association, 1994). Secondly, several self, parent and teacher rating scales for measuring symptoms of psychiatric disorders exist and are recommended for assessment and treatment validation. Often these instruments are used for screening of mental disorders both in community and clinical settings. Lastly, standardized observation protocols and self monitoring instruments exist as well. However, these methods are time consuming and costly and, thus, are of limited use in clinical settings (Pelham et al., 2005). Despite the use of these standardized instruments and the corresponding improvement of diagnostic evidence, some limitations remain and are addressed in the present literature (Gray, 2004). For example, it was argued that further evidence is needed regarding the analytic procedure to integrate diagnostic information from several instruments and sources.

2.1 Psychiatric diagnoses

The commonly used classification systems for psychiatric disorders are the International Classification of Diseases (ICD) and the Diagnostic and Statistical Manual of Mental Disorders (DSM). These classification systems allow clinicians and researchers from around the world to identify psychiatric disorders according to standardized criteria and to communicate about them.

The ICD-10 was endorsed by the Forty-third World Health Assembly in May 1990 and came into use in the World Health Organization (WHO) Member States as from 1994 (World Health Organization, 1994). The ICD-10 is the international standard of diagnostic classification for all general epidemiological, many health management purposes and for clinical use. The section F addresses psychiatric disorders. (World Health Organization, 1994).

The DSM-IV is as the ICD-10 a categorical classification system. The DSM-IV is published by the American Psychiatric Association and provides diagnostic criteria for mental disorders. It is used in the United States and in varying degrees around the world by clinicians, researchers, psychiatric drug regulation agencies, health insurance companies, pharmaceutical companies and by policy-makers (American Psychiatric Association, 1994).

Both classification systems include diagnostic criteria that were specified to evaluate mental health problems in children and adolescents. Although ICD-10 and DSM-IV criteria are not identical for most psychiatric disorders, the majorities of the diagnoses are comparable. The diagnostic criteria of most psychiatric disorders include the age of onset, the duration, the frequency and the severity of the symptoms and the impairment of these symptoms. In the present studies either ICD-10 or DSM-IV criteria were applied.

The present thesis deals with the prediction of three common psychiatric disorders in children and adolescents: ADHD, clinical depression and ODD. A short description of these disorders is given below.

2.1.1 Attention-Deficit-Hyperactivity Disorder (ADHD)

Attention-Deficit Hyperactivity Disorder (ADHD; DSM IV 314.01) or hyperkinetic disorder (ICD-10 F90.0) is a psychiatric disorder which begins in early childhood. It affects about 3 to 5% of school children (Polanczyk, de Lima, Horta, Biederman, & Rohde, 2007) with symptoms starting before the age of seven. It is characterized by a persistent pattern of impulsiveness and inattention, with or without a component of hyperactivity. ADHD occurs twice as commonly in boys as in girls (Dulcan & Benson, 1997). ADHD is generally a chronic disorder with 10 to 40% of individuals diagnosed in childhood continuing to meet diagnostic criteria in adulthood. As they mature, adolescents and adults with ADHD are likely to develop coping mechanisms to compensate for their impairment.

2.1.2 Clinical depression

Major depressive disorder (DSM-IV, 296.2/296.3) or depressive episode (ICD-10, F32/F33) is a mental disorder typically characterized by a pervasive low mood, low self-esteem and loss of interest or pleasure in usual activities. The general term depression is often used to describe the disorder, but since it is also used to describe temporary sadness or a depressed mood, more precise terminology is preferred in clinical use and research. Major depression is an often disabling condition which adversely affects a person's family, work or school life, sleeping and eating habits and general health.

A child with depression may pretend to be sick, refuse to go to school, cling to a parent or worry that a parent may die. Older children may sulk, get into trouble at school, be negative and irritable and feel misunderstood. Before puberty, boys and girls are equally likely to

develop depressive disorders. By age 15, however, girls are twice as likely as boys to have experienced a major depressive episode (Cyranowski, Frank, Young, & Shear, 2000).

2.1.3 Oppositional defiant disorder (ODD)

Oppositional defiant disorder (ODD; DSM-IV 313.81, ICD-10 F91.3) is a negativistic pattern of hostile and defiant behavior that has been present for at least 6 month. ODD is frequently associated with ADHD and other psychiatric disorders, in particular with conduct disorders (CD). Untreated, about 52% of the children with ODD will continue to meet the DSM-IV criteria up to three years later and about half of those 52% will progress into a CD (Lahey, Loeber, Quay, Frick, & Grimm, 1992). It was argued that CD is a more severe form of ODD (Loeber, Keenan, Lahey, Green, & Thomas, 1993). However, recent studies have identified ODD as a separate disorder from CD according to comorbidity and impairment (Greene et al., 2002).

2.2 Diagnostic interviews

A diagnostic assessment may be defined as gathering information to estimate the likelihood of various diagnostic probabilities. At the beginning of the assessment, the clinician has an initial impression of the child and adolescent problems. This can be described as a pretest probability of a psychiatric disorder (Richardson, Wilson, & Guyatt, 2002). After the diagnostic assessment, the likelihood of a given diagnosis after appropriate tests have been conducted can be formulated. This diagnostic process from a pretest probability to a posttest probability can include various forms of structured or unstructured interviews and diagnostic tests (Doss, 2005).

In child and adolescent psychiatry, the diagnostic process typically takes the form of an unstructured interview in which clinicians follow up on their initial diagnostic impression by asking questions to rule in or out diagnoses (Doss, 2005). However, this unstructured

method is limited regarding reliability and validity. Accordingly, Angold (2002) has suggested that several biases associated with unstructured clinical decisions may affect the validity of the corresponding diagnoses. First, clinicians may have a tendency toward making a decision before all information has been collected. Secondly, diagnostic decisions may be taken regarding problems that are familiar to the clinician. Other less prominent problems may not be recognized and therefore be neglected by the clinician. Thirdly, clinicians may be interested to assign one diagnosis that may be required for administrative purposes but time constraints and workload prohibit a comprehensive assessment to assign multiple diagnoses (A. L. Jensen & Weisz, 2002).

An alternative approach to assign psychiatric diagnoses is the use of standardized interviews. Various instruments for parents and adolescents in different languages are available (e.g. Diagnostic Interview Schedule for Children's, DISC; Parental Account of Childrens Symptoms, PACS, Chen & Taylor, 2006; Shaffer et al., 1996). Although structured and semi-structured interviews are often used in research settings and show a reliable measure of psychiatric disorders, their use in clinical settings is limited for the following reasons (A. L. Jensen & Weisz, 2002). First, the major purpose of interviewing patients or their parents is often to identify issues that need to be addressed in treatment. Clinicians may be focused on primary problems rather than on secondary problems that may not be part of the treatment agenda. Secondly, interviewing is done by heavily scheduled clinical staff and may be more based on own expertise than strictly on ICD-10 or DSM-IV criteria (Robins, 2002). Thirdly, time pressures brought on by cost and productivity policies limit time that can be devoted to clinical interviews. Fourthly, given the size and the complexity of the diagnostic classification systems that are relevant for the age group of children and adolescents, a structured interview is almost certain to be quite lengthy and expensive.

From an evidence-based point of view, structured interviews are superior regarding objectivity and reliability. However, structured interviews are costly and time-consuming and therefore of limited use in clinical settings. Recently, internet-based forms of diagnostic

interviews have been developed. The “Development and Well-Being Assessment” (DAWBA; Goodman, Ford, Richards, Gatward, & Meltzer, 2000) is a novel package of questionnaires, interviews, and rating techniques designed to generate ICD-10 and DSM-IV psychiatric diagnoses on 5 to 16-year-olds by adolescent, parent and teacher information. DAWBA can be filled out and administered online (http://www.dawba.com) and is therefore much easier to handle as conventional forms of structured or semi-structured interviews.

However, these online instruments are quite novel and not yet widespread. An alternative to conventional diagnostic interviews are the use of standardized questionnaires and rating scales in clinical assessments. Advantages and disadvantages of these instruments are presented in the following chapter.

2.3 Rating scales

Parent and self rating scales are often used in clinical assessments for screening purposes. Furthermore, specific symptom scales are helpful for the diagnostic decision process and to quantify the severity of the symptoms. In addition, rating scales were inserted for treatment evaluation.

Diagnostic rating scales are mostly used as paper-and-pencil tests. Adolescents, parents and teachers are requested to refer to a given statement (e.g. “I feel sad”) by selecting a category label from a list indicating the extent of disagreement or agreement with a statement (e.g. strongly disagree). These categories are quantified by different values (also known as Likert scale) allowing to calculate averages and further arithmetic operations. Furthermore, a total score of a scale can be built by summarising values of related items. Further information about scales construction and item analyses is given in the present handbooks of test construction (e.g. DeVellis, 2003).

Psychometric properties can be applied to rating scales. The key traditional concepts in classical test theory are reliability and validity. A reliable measure is measuring something consistently while a valid measure is measuring what it is supposed to be measured. A reliable measure may be consistent without necessarily being valid. As a measure of reliability, the internal consistency of a rating scale can be described by the Cronbach alpha measure (Gray, 2004) (Cronbach, 1951).

Self and parental rating scales usually based on normative samples and cut-off scores, are recommended according to a T-score. This T-score is based on 1 to 2 standard deviations above the average score taking into account the severity and the frequency of the symptoms described. However, the prevalence of psychiatric disorders varies and more than the identified fixed percentage of children above the defined norm can suffer from the symptoms. Despite the normative evaluation, rating scales are of limited use for identifying psychiatric diagnoses because they do not include information about other diagnostic criteria as the onset, the durance and the impact of the symptoms.

Hence, apart from using T-scores to evaluate the severity and the count of the symptoms compared to a normative sample, behavior checklist should be tested for their accuracy to predict psychiatric disorders.

3 Diagnostic accuracy of rating scales

Testing diagnostic accuracy of a certain rating scale is complex and requires specific methodological knowledge. Exaggerated and biased results of studies testing diagnostic accuracy can lead the examiner into making incorrect treatment decisions. Therefore, the STARD initiative (Bossuyt et al., 2003) aimed to improve the accuracy and completeness of studies dealing with diagnostic accuracy. It provided a 25-item checklist as a guideline for studies dealing with diagnostic accuracy (see Appendix 9.1). Most of these items are addressed in the present chapter.

When testing the diagnostic accuracy of a rating scale, the predictive and the concurrent validity is addressed. Various methods to define diagnostic accuracy with different statistical methods exist and will be discussed in this chapter. The accuracy of a diagnostic test is generally assessed in a cross-validation study in which subjects are evaluated with both a diagnostic "gold standard" measure and the diagnostic test under evaluation. Several important issues can affect the validity of a diagnostic test but the choice of the "gold standard" and the choice of the subjects are the two major issues (Gray, 2004).

3.1 The "gold standard" of a diagnostic test

The diagnostic assessment process in order to come to an ICD-10 or DSM-IV diagnoses was addressed in a previous chapter. However, for the definition of a diagnostic "gold standard" is an issue of particular importance. In some branches of medicine the "gold standard" may refer to one or more laboratory tests with unambiguous results. In child and adolescent psychiatry no laboratory test exists for such diagnoses. Diagnostic criteria refer to the mental status and behavior of the subject. Thus, the information needed is based on the phenomenological description of the subject itself or of behavioral ratings of others. However, there are unresolved issues concerning the validity of the diagnostic criteria (e.g. see the discussion for DSM-V or ICD-11 criteria in Kupfer, Regier, & Kuhl, 2008; Regier, 2007).

Furthermore, there are methodological limitations in order to elicit symptoms and to use this information to arrive at a diagnosis. Despite the good reliability of standardized instruments such as structured interviews, the question of validity remains unsolved. Furthermore, due to limited cognitive development, statements of young children are often ambiguous and conflicting. Thus, diagnostic interviews are available for 11-years-olds and older youth but not for younger children. In general, parent interviews are used to assess mental problems of younger children. However, this information may be biased by the parents' experience with the child. Thus, aggressive or oppositional behavior involving parents can be overestimated whereas other less prominent emotional problems may be ignored by the parents. In studies with adolescents and their parents, small rates of agreements between mental health problems of the adolescent were found (e.g. P. S. Jensen, Salzberg, Richters, & Watanabe, 1993). Until now, no guidance has been provided how to deal with disagreeing information of different informants in order to come to DSM-IV or ICD-10 psychiatric disorders. These limitations concerning the validity of DSM-IV and ICD-10 diagnosis have to be taken into account when deciding for a "gold standard" of psychiatric disorder.

For the following first and the third study of the present thesis, the "gold standard" is based on a structured diagnostic interview with a parent when describing the criteria for ADHD and ODD. As most of the ADHD and ODD diagnostic criteria are based on observable behavior rather than on mental state processes, this approach seems adequate. In the second study, dealing with adolescent depression, another "gold standard" was applied. Despite the problems concerning reliability, clinical diagnosis based on a best estimate procedure was used. This procedure has the advantage that full information of various instruments and informants were included in order to come to the final diagnoses.

3.2 Choice of subjects

A cross-sectional study including subjects, similar to those to whom the rating scale is expected to be administered in clinical practice, is the most appropriate design (Gray, 2004).

However, most rating scales were used in different settings, such as in community samples for screening and research purposes and in clinical settings for treatment planning. Few rating scales were tested in both settings for the prediction of a target diagnosis. When a rating scale is tested in a mixed sample including very ill patients from a clinically referred sample and healthy controls, the rating scale will perform better in distinguishing the ill from the healthy than in actual practice. This bias has been labeled as “spectrum bias” (Knotterus, 2002).

In addition, case control studies were applied for testing diagnostic accuracy. In case control studies a sample of patients is compared to a normative community sample. Case control designs have a more exploratory character because the tested sample is not representative for the population in which the test should be used (Sullivan Pepe, 2003).

3.3 Diagnostic accuracy of a dichotomous test

A variety of terms are used to describe the performance of a diagnostic test. First, a test result is to be considered to be dichotomous. The following section deals with the diagnostic accuracy of rating scales providing a continuous measure. In both sections, the absence or presence of the disorder is based on the “gold standard” for assessing psychiatric disorders. Therefore, the following methods for describing diagnostic accuracy are limited to the validity of the “gold standard” measure.

3.3.1 The basic model

When both the “gold standard” and the evaluating test are positive, the result of the diagnostic test is considered to be true positive (table 1, TP). Likewise, if both yield negative results the test is considered to be true negative (table 1, TN). If the test gives a positive test result but the “gold standard” is negative, the diagnostic test is to be considered false positive (table 1, FP). Vice versa, if the test result is negative and the “gold standard” is positive, the

result of the test is false positive (table 1, TP). The addition of the TP, TN, FP and FN rates is considered to be 1 and includes the entire sample of subjects which was tested.

Table 1. *Possible results of a dichotomous diagnostic test*

	Disorder present	**Disorder absent**	**Totals**
Test Result positive	True positives (TP)	False positives (FP)	TP+FP
Test result negative	False negative (FN)	True negative (TN)	FN+TN
Totals	TP+FN	FP+TN	TP+FP+FN+TN = 1

3.3.2 Sensitivity, Specificity, positive predicted value and negative predicted value

From the rate of TP, TN, FP and FN, key values for describing the diagnostic performance of a test can be computed. Accordingly, sensitivity (SE) refers to the proportion of subjects with the disorder (as assessed by the "gold standard") who are detected by the diagnostic test. A highly sensitive test will detect most of the cases with the disorder. Likewise, the specificity (SP) of a diagnostic test is the proportion of the subjects without the disorder according to the "gold standard" which is found negative by the diagnostic test. A highly specific test will not misidentify healthy subjects as having a disorder in terms of the probabilities used in table 1: SE = TP / (TP + FN), and SP = TN / (FP + TN). Both of these measures can range from 0 to 1. Although it seems counterintuitive, highly sensitive diagnostic tests are most useful to rule out diagnoses whereas highly specific tests are most suited to rule in diagnoses. However, when describing a diagnostic test both specificity and sensitivity have to be reported as these two measures are reciprocally related. One indicator alone is not sufficient as a measure of diagnostic accuracy.

Further measures of diagnostic accuracy are the positive and negative predictive values. The positive predictive value (PPV) represents the probability that an individual with a positive test result really has the diagnosis. In contrast, the negative predictive power (NPV) refers to the probability that an individual with a negative test result does not have the diagnosis. In

contrast to SE and SP, both PPV and NPV are dependent on the prevalence of the disorder in the tested sample.

3.3.3 Likelihood ratios

As an alternative method to estimate whether or not a specific individual has the diagnosis after the presence of a test result, likelihood ratios (LR) can be calculated. LR are independent of the prevalence of the disorder in the tested sample. The likelihood ratio of a positive test result (LR+) is the ratio of the likelihood (probability) of a positive test result in the population of the diagnosed subjects and the likelihood of a positive test result in the population of non-diagnosed subjects. Similarly, the likelihood ratio of a negative test result (LR-) is the rate of the likelihood of a negative test result in the population of the diagnosed subjects and the likelihood of a negative test result in the population of non-diagnosed subjects. LR+ and LR- can be calculated from TP, TN, FP and FN. In addition, likelihood ratios can be calculated directly from sensitivities and specificities:

$$LR+ = [TP / (TP + FN)] / [FP / (FP + TP)] = SE / (1 - SP)$$

$$LR- = [FN / (TP + FN)] / [TP / (FP + TP)] = (1 - SE) / SP$$

Likelihood ratios can be treated like odds ratios. They are an intuitive measure in regards to the chances to have the diagnoses after the reception of a positive test or negative test result. A LR greater than 1 indicates that the test result is associated with the presence of the disorder whereas a likelihood ratio less than 1 indicates that the test result is associated with the absence of a disorder. Furthermore, LR can be used to estimate posttest odds if pretest odds are known by the following formula: posttest odds = pretest odds x LR (Deeks & Altman, 2004).

3.3.4 Efficiency

Finally, the efficiency (EFF) of a diagnostic test is defined as the probability that the test and the diagnoses match. Thus, EFF is the sum from TP and TN in the tested sample (EFF = TP + TN). Like SE, SP and PPV and NPV, EFF is an uncalibrated measure with a random value that depends on the prevalence of the disorder and the level of a test. The level of a test (Q) is defined by Q = TP + FN and can be described as the general probability of a test to receive a positive test result (Kraemer, 1992). Kraemer (1992) therefore proposes a quality coefficient of efficiency correcting for the independence of the prevalence (P) in the sample and to take into account the rate of a positive test result (Q). A quality index of efficiency can be calculated using the following formula: $d_Q = [EFF – PQ - (1 - P)(1 - Q)]/[1 - PQ- (1 - P)(1 - Q)]$.

Apart form the indicators above, further measures of diagnostic accuracy as diagnostic odds ratios and error rates have been described. As these indicators were not used in the following studies, they are not presented here. Interested readers are referred to the statistic literature (e.g. Sullivan Pepe, 2003)

3.4 Diagnostic accuracy of a rating scale (continuous measure)

Rating scales generally have more than just two values. A clinical rating scale produces a score which is based on the count of symptoms that is present in a given time. Thus, rating scales are reflecting dimensional qualities of a child's emotional and behavioral problems whereas a psychiatric diagnosis is reflecting a categorical approach to child's emotional and behavioral problems. However, often categorical psychiatric diagnoses are needed for medical decision making (e.g. treatment planning). Thus, tests for diagnostic accuracy of a rating scale to predict psychiatric diagnoses and subsequent cut-off analyses were recommended. These methods are described in the present section. Furthermore, statistical methods for comparing various rating scales and multidimensional modes which include two or more rating scales for predicting a disorder are presented.

3.4.1 Receiver operating characteristic analysis

Receiver operating characteristic (ROC) analysis is a widely used and accepted method for improving decision making performance across a range of diagnostic settings. ROC analyses were first developed in signal detection theory for assessing the predictive value of a test for a "gold standard". Since its beginning in the 1950s (described in Green & Swets, 1966), ROC techniques were inserted in various scientific branches. For example, ROC analyses were used in weather forecasting (e.g. Mason, 1982), aptitude testing (Stillman & Duncan, 2005), medical imaging (for an overview see Obuchowski, 2003) and in general, for medical decision making. For an overview of ROC analyses and its significance for various areas, a very descriptive summary has been presented by Swets (1988).

As an expansion of the "basic model" (see 3.3.1), sensitivities and specificities can be computed for all possible cut-off scores of a rating scale. A ROC curve plots the sensitivity against the (1 – specificity) for each possible cut-off value. Therefore, a ROC curve can visually demonstrate the cut-off scores that efficiently maximize both sensitivity and specificity. The point nearest the upper left corner on a ROC curve shows the best abilities according to sensitivity and specificity (figure 1). The most common index of accuracy in ROC analysis is the area under the curve (AUC) which assesses the possibility of correctly classifying a randomly selected pair of subjects in which one is a case and one is a non-case. AUC values range between .5 in which correct classification occurs in 50% of the cases, and 1.0 in which correct classification occurs with every case (sensitivity and specificity = 1). Acceptable AUC values vary depending on the base rate of the diagnosis and other sample characteristics. AUCs as a measure of excellence for predicting diagnosis should be interpreted as follows: poor (50-.70); moderate to fair (.70-.80); good (.80-.90) and excellent (.90-1.00) (Ferdinand, 2008).

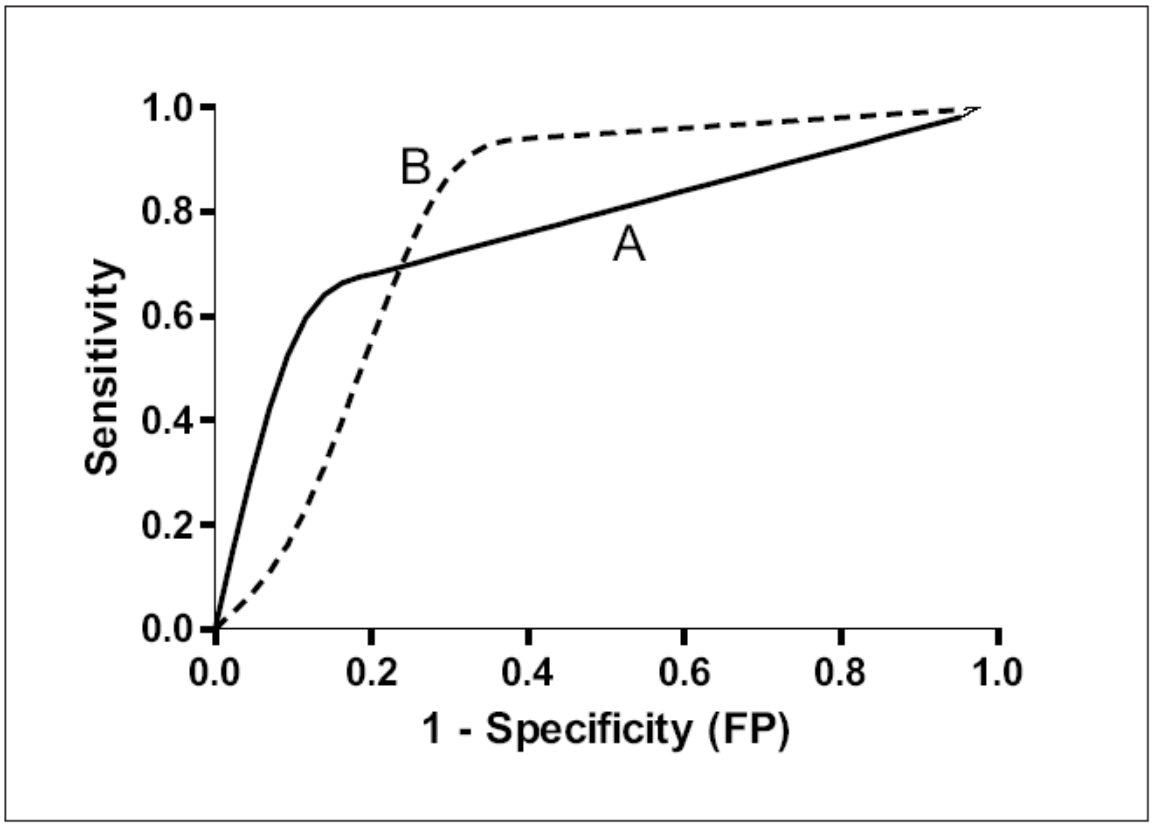

Figure 1 Two different ROC curves with the same area under the curve but different shapes

Figure 1 shows two possible ROC curves with the same AUC but different shape. Scale A (ROC curve A) is better suited to rule out diagnosis as most of the possible cut-off scores have high sensitivities. Otherwise, scale B (ROC curve B) is better in ruling in diagnosis as most of the possible cut-off scores have high specificities. However, the AUC is equal for both ROC curves, therefore both tests show a comparable general diagnostic accuracy.

As seen before, the AUC can be described as a general measure of the diagnostic accuracy of a rating scale. Thus, the AUC is mostly used for estimating the predictive power of a rating scale. Most statistic programs include a feature to calculate ROC curves and AUC measures (e.g. SPSS inc., 2006). In addition, several computer programs for ROC analyses are available, most can be downloaded from the Internet (Stephan, Wesseling, Schink, & Jung, 2003). For calculating AUC, several mathematical methods exist (Hanley & McNeil, 1982; Zhou, 1996) but mostly a maximum likelihood estimation has been used. For comparison of

different AUCs within the same sample, a critical z-ratio can be calculated using a formula correcting for the non-independence of the scales (Hanley & McNeil, 1983).

3.4.2 Cut-off score analyses

As seen in the previous section, visual ROC graphs can be used to show various cut-off scores of a scale. A cut-off score can be directly derived from the ROC curve, the point nearest the upper left corner shows the perfect balance of maximizing sensitivity and specificity. This point which can be geometrically determined is also the cut-off score showing the highest efficiency (EFF, see 3.3.4). Although this method has been used frequently in child and adolescent psychiatric studies (e. g. Christiansen et al., 2008; Lampert, Polanczyk, Tramontina, Mardini, & Rohde, 2004), the results are of limited validity. The magnitude of EFF depends strongly on the level of a test and its relation to the prevalence of the disorder in the sample (Kraemer, 1992). In other words, EFF has to be calibrated in order to come to valid results which can be compared to other studies of diagnostic accuracy with different base rates. These calculations were shown in a previous section of this chapter (see 3.3.4).

However, even if a quality measure like quality efficiency was used, the resulting cut-off scores are not always useful for clinical practice. First, as mentioned before next to rating scales also other methods are used in diagnostic assessments. In order to come to a final diagnosis, rating scales are often used as initial screening instruments. Therefore, rating scales should maximize sensitivity to include probable subjects. Secondly, different costs and benefits of patients with false positive and false negative disorders have to be taken into account. Under the consideration of medical consequences of errors and the costs of a test, corrected efficiency measures as proposed by Kraemer (1992) should be applied.

3.4.3 Further statistical methods for testing and comparing different rating scales

Often other statistical analyses were performed in combination with ROC analyses and cut-off score analyses. If more than one rating scale is needed to be tested to predict a specific diagnosis, the AUC's of these scales can be compared by the use of a critical z-test. Furthermore, on a visual presentation of the ROC curve one or another test can be preferred by interpreting the shape and form of the curve. However, if a combination of scales and measures are needed to be tested in their ability to discriminate disordered from non-disordered youth, logistic regression analyses or discriminant analyses have to be used. Both of these methods provide an indicator on which it can be evaluated how good a specific model consisting of multiple rating scales is able to predict a dichotomous result (e.g. disorder vs. non-disorder). In addition, logistic regression analyses provide overall probabilities based on the tested multivariate model that can be used in subsequent ROC analyses.

3.5 References

American Academy of Child and Adolescent Psychiatry. (2005). *Evidence Based Practice*, from http://www.aacap.org/cs/root/policy_statements/evidence_based_practice

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC: Author.

Angold, A. (2002). Diagnostic interviews with parent and childrens. In M. Rutter & E. Taylor (Eds.), *Child and Adolscent psychiatry: Modern Approaches, 4th edition* (pp. 32-51). Oxford: Blackwell Scientific.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Bmj, 326*(7379), 41-44.

Chen, W., & Taylor, E. (2006). Parental account of children's symptoms (PACS), ADHD phenotypes and it's application to melecular genetic studies. In R. D. Oades (Ed.),

Attention-deficit/hyperactivity disorder and the hyperkinetic syndrome: current ideas and ways forward (pp. 3-20). Hauppauge NY: Nova Science Publishing Inc.

Christiansen, H., Chen, W., Oades, R. D., Asherson, P., Taylor, E. A., Lasky-Su, J., et al. (2008). Co-transmission of conduct problems with attention-deficit/hyperactivity disorder: familial evidence for a distinct disorder. *Journal of Neural Transmission, 115*(2), 163-175.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-333.

Cyranowski, J. M., Frank, E., Young, E., & Shear, M. K. (2000). Adolescent onset of the gender difference in lifetime rates of major depression: a theoretical model. *Arch Gen Psychiatry, 57*(1), 21-27.

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *Bmj, 329*(7458), 168-169.

DeVellis, R. F. (2003). *Scale Development: Theories and Applications*. Thousend Oaks: Sage Publications.

Doss, A. J. (2005). Evidence-based diagnosis: incorporating diagnostic instruments into clinical practice. *J Am Acad Child Adolesc Psychiatry, 44*, 947-952.

Dubicka, B., & Wilkinson, P. (2007). Evidence-based treatment of adolescent major depression. *Evid Based Ment Health, 10*(4), 100-102.

Dulcan, M. K., & Benson, R. S. (1997). AACAP Official Action. Summary of the practice parameters for the assessment and treatment of children, adolescents, and adults with ADHD. *J Am Acad Child Adolesc Psychiatry, 36*(9), 1311-1317.

Evans, S. W., & Youngstrom, E. (2006). Evidence-based assessment of attention-deficit/hyperactivity disorder: measuring outcomes. *J Am Acad Child Adolesc Psychiatry, 45*(9), 1132-1137.

Ferdinand, R. F. (2008). Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. *J Anxiety Disord, 22*(1), 126-134.

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*(5), 645-655.

Gray, G. E. (2004). *Evidence Based Psychiatry*. Washington DC: American Psychiatric Publishing.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greene, R. W., Biederman, J., Zerwas, S., Monuteaux, M. C., Goring, J. C., & Faraone, S. V. (2002). Psychiatric comorbidity, family dysfunction, and social impairment in referred youth with oppositional defiant disorder. *American Journal of Psychiatry, 159*(7), 1214-1224.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839-843.

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *J Consult Clin Psychol, 70*(1), 158-168.

Jensen, P. S., Salzberg, A. D., Richters, J. E., & Watanabe, H. K. (1993). Scales, diagnoses, and child psychopathology: I. CBCL and DISC relationships. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 397-406.

Knotterus, J. A. (2002). *The Evidence Base of Clinical Diagnosis*. London: BMJ Books.

Kraemer, H. C. (1992). *Evaluating medical tests. Objective and quantitative guidelines.* Newbury Park: Sage Publications, Inc.

Kupfer, D. J., Regier, D. A., & Kuhl, E. A. (2008). On the road to DSM-V and ICD-11. *Eur Arch Psychiatry Clin Neurosci, 258 Suppl 5*, 2-6.

Lahey, B. B., Loeber, R., Quay, H. C., Frick, P. J., & Grimm, J. (1992). Oppositional defiant and conduct disorders: issues to be resolved for DSM-IV. *J Am Acad Child Adolesc Psychiatry, 31*(3), 539-546.

Lampert, T. L., Polanczyk, G., Tramontina, S., Mardini, V., & Rohde, L. A. (2004). Diagnostic performance of the CBCL-Attention Problem Scale as a screening measure in a sample of Brazilian children with ADHD. *J Atten Disord, 8*(2), 63-71.

Loeber, R., Keenan, K., Lahey, B. B., Green, S. M., & Thomas, C. (1993). Evidence for developmentally based diagnoses of oppositional defiant disorder and conduct disorder. *J Abnorm Child Psychol, 21*(4), 377-410.

Mason, I. (1982). *Australian Meteorology, 30*(291).

Nair, J., Ehimare, U., Beitman, B. D., Nair, S. S., & Lavin, A. (2006). Clinical review: evidence-based diagnosis and treatment of ADHD in children. *Mo Med, 103*(6), 617-621.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology, 229*(1), 3-8.

Pelham, W. E., Jr., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *J Clin Child Adolesc Psychol, 34*(3), 449-476.

Polanczyk, G., de Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *Am J Psychiatry, 164*(6), 942-948.

Regier, D. A. (2007). Dimensional approaches to psychiatric classification: refining the research agenda for DSM-V: an introduction. *Int J Methods Psychiatr Res, 16 Suppl 1 2007*, S1-5.

Reichow, B., Volkmar, F. R., & Cicchetti, D. V. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *J Autism Dev Disord, 38*(7), 1311-1319.

Richardson, W. S., Wilson, M., & Guyatt, G. (2002). The diagnostic process. In G. Guyatt & D. Rennie (Eds.), *User's Guide to the Medical Literature*. Chicago: AMA.

Robins, L. N. (2002). Birth and develpment of psychiatric interviews. In M. T. Tsuang & M. Thohen (Eds.), *Textbook of Psychiatric Epidemiology*. New York: Wiley-Liss.

Shaffer, D., Fisher, P., Dulcan, M. K., Davies, M., Piacentini, J., Schwab-Stone, M. E., et al. (1996). The NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA Study. Methods for the Epidemiology of Child and Adolescent Mental Disorders Study. *J Am Acad Child Adolesc Psychiatry, 35*(7), 865-877.

Silverman, W. K., & Ollendick, T. H. (2005). Evidence-based assessment of anxiety and its disorders in children and adolescents. *J Clin Child Adolesc Psychol, 34*(3), 380-411.

SPSS inc. (2006). Statistical packet for social scientists, SPSS. Illinois.

Stephan, C., Wesseling, S., Schink, T., & Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem, 49*(3), 433-439.

Stillman, J. A., & Duncan, J. R. J. (2005). A detection theory approach to the evaluation of assessors in assessment centres. *Journal of Occupational and Organizational Psychology, 78*, 581-594.

Sullivan Pepe, M. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: University Press.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

World Health Organization. (1994). *The ICD-10 classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines*. Geneva, Switzerland: World Health Organization.

Zhou, X. H. (1996). Empirical Bayes combination of estimated areas under ROC curves using estimating equations. *Med Decis Making, 16*(1), 24-28.

4 Study 1: Accuracy of the DSM-oriented attention problem scale of the Child Behavior Checklist in diagnosing Attention-Deficit-Hyperactivity Disorder[1]

4.1 Abstract

Objective: The present study aimed at testing the child behavior checklist (CBCL) including an adapted 5-item DSM-oriented attention problem scale for predicting attention deficit hyperactivity disorders (ADHD). Methods: CBCL ratings were made both in a community sample (N = 390) and an outpatient child psychiatric sample (N = 392). Four different prediction models were analyzed in a community sub-sample (n = 195) and an outpatient sub-sample (n = 196) and cross-validated in two further sub-samples of the same size. Results: The adapted DSM-oriented attention problem scale was superior to the original attention problem scale in the identification of ADHD subjects. A raw score of 5 to 6 on the reduced DSM-oriented attention problem scale was the best discriminator between cases and non-cases. Conclusions: The adapted DSM-oriented attention problem scale of the CBCL is a useful screening instrument for ADHD with adequate diagnostic accuracy in community and outpatient samples.

Keywords: ADHD, attention problems, Child Behavior Checklist, prediction

4.2 Introduction

There are various ways to obtain diagnostic information on attention problems in children and adolescents. In clinical assessments, semi-structured or structured interviews are mainly

---

[1] Aebi, M., Winkler Metzke, C. & Steinhausen H.-Ch. (in press). Accuracy of the DSM-oriented attention problem scale of the Child Behavior Checklist in diagnosing Attention-Deficit-Hyperactivity Disorder. *Journal of Attention Disorders* DOI:

used for diagnosis. These methods require trained staff and are time-consuming and expensive. On the other hand, behavior checklists are easy to use and have proven to be efficient and low cost measures for the identification of behavioral and emotional problems in children and adolescents. However, behavior checklists do not provide psychiatric diagnoses. Advantages and disadvantages of both approaches have been discussed extensively with no clear preference of either method (Ferdinand et al., 2004; Kraemer, Noda, & O'Hara, 2004).

The Child Behavior Checklist (CBCL) is a worldwide used parental questionnaire that reports emotional and behavioral difficulties in children and adolescents. Several studies have demonstrated convergence between CBCL scales and various disorders in community and clinic-referred samples (Edelbrock & Costello, 1988; Ferdinand et al., 2004; Kazdin & Heidish, 1984; Steinhausen, Winkler Metzke, Meier, & Kannenberg, 1997). However, other studies have been less supportive. For instance, in a community based study Jensen, Salzberg, Richters, & Watanabe (1993) found that the CBCL showed only modest ability in predicting psychiatric diagnoses.

Further studies have been specifically addressing the prediction of interview-based Attention-Deficit-Hyperactivity Disorder (ADHD) diagnosis by the CBCL scales (Biederman et al., 1993; Chen, Faraone, Biederman, & Tsuang, 1994; Doyle, Ostrander, Skare, Crosby, & August, 1997; Eiraldi, Power, Karustis, & Goldstein, 2000; Hudziak, Copeland, Stanger, & Wadsworth, 2004; Lampert, Polanczyk, Tramontina, Mardini, & Rohde, 2004; Ostrander, Weinfurt, Yarnold, & August, 1998; Steingard, Biederman, Doyle, & Sprich-Buckminster, 1992; Zelko, 1991) and have reported rather conflicting findings. Whereas some of these studies revealed evidence that the attention problem scale predicts ADHD (Biederman et al., 1993; Chen et al., 1994; Eiraldi et al., 2000; Hudziak et al., 2004; Steingard et al., 1992; Zelko, 1991) other studies failed to confirm these results (Doyle et al., 1997; Ostrander et al., 1998). The latter studies reported that the social problem scale or the aggressive behavior scale of the CBCL were even stronger related to ADHD than the attention problem scale. In

some studies diagnostic accuracy was only moderate or insufficient as indicated by a low predictive power of the attention problem scale or a poor potential to rule out non-ADHD subjects (Doyle et al., 1997; Eiraldi et al., 2000; Lampert et al., 2004). However, the attention problem scale has also been reported to discriminate sufficiently between subjects with ADHD and unreferred controls in a study by Biederman and colleagues (1993). These authors found an excellent odds ratio of 99.1 and a total predictive power of .86 of the attention problem scale using a cut-off score of T = 60. The accurate performance of the attention problem scale has been also confirmed by Chen et al. (1994) in a sample of subjects with ADHD compared to pediatric controls and by Hudziak et al. (2004) in a mixed sample of community and outpatient subjects with high or low levels of behavioral and attentional problems and their siblings. Both studies found that an even lower cut-off score of T = 55 was more efficient in diagnostic assessment (sensitivity .61-.88., specificity . 83 -.94). However, generalizability of these results may be limited due to a selection bias because a scale performs much better in the comparison of severely impaired patients with healthy controls than in clinical samples (Gray, 2004).

Various reasons may account for these inconsistent findings. First, the very common comorbid disorders in ADHD subjects could have exerted an uncontrolled impact on assessment. Accordingly, the attention problem scale has been shown to be highly correlated with other CBCL scales including the social problem scale and the aggressive behavior scale (Doyle et al., 1997; Eiraldi et al., 2000; Jensen et al., 1993). Secondly, different univariate and multivariate methods have been applied in order to test diagnostic accuracy of the CBCL. Therefore, results may have been influenced by different methodological approaches. Thirdly, sample effects may also have had an impact on the results. The attention problem scale has been superior in predicting ADHD if the sample included different recruitment sources and therefore a broad spectrum of referred and non-referred subjects (Chen et al., 1994; Hudziak et al., 2004). In homogenous samples, results have been less clear (Doyle et al., 1997; Eiraldi et al., 2000; Jensen et al., 1993). Fourthly, the original construction of the CBCL problem scales did not refer to the categorical

approach of ICD or DSM diagnostic criteria. On the contrary, the scales had been empirically defined and reflect a dimensional approach to psychopathology. Although the CBCL attention problem scale bears an a priori resemblance to ADHD, the scale is also including items which are not related to ADHD criteria (e.g., items # 13. confused or # 61. poor school work).

In order to overcome the differences in the theoretical approach, DSM-oriented scales have been introduced in the 2001 revision of the CBCL and have been recommended both for clinical and research purposes. These new scales are based on the original CBCL item pool by expert ratings of similarity to DSM-IV criteria and show good psychometric properties (Achenbach, Dumenci, & Rescorla, 2003). The DSM-oriented attention-deficit-hyperactivity problem (DSM-ADH) scale includes 7 items and has been found to correlate significantly with clinical DSM-IV diagnoses (Achenbach & Rescorla, 2001). However, so far the potential of the DSM ADH scale in the prediction of ADHD has not yet been tested. In the present study, an adapted 5-item DSM-ADH-scale based on the 1991 edition of the CBCL was used. It seemed worthwhile to perform these analyses because the 1991 edition of the CBCL still is widely used outside the US.

Thus, the present study aimed at testing the diagnostic accuracy of the adapted CBCL DSM-ADH-scale compared to the original attention problem scale. In order to overcome sampling effects due to different kinds of recruitment, we studied this question separately in a referred psychiatric sample and in a community based sample with homogeneous recruitment. These different samples were chosen because the CBCL can either be used as part of the clinical diagnostic assessment or as a screening device in community surveys in order to identify subjects at risk of major psychopathology.

Furthermore, in order not to obtain findings that reflect only variations in the sample rather than variations in the population, the psychiatric and the community based samples were each split into two sub-samples so that the results could be cross-validated. Clearly, cross-

validation and generalizability tests of the results are mandatory (Leon, Olfson, Weissman, Portera, & Sheehan, 1996). So far, only two studies dealing with the prediction of ADHD by the CBCL (Chen et al., 1994; Hudziak et al., 2004) used cross-validation in order to improve reliability and validity of the results.

Finally, due to the frequent comorbid disorders in ADHD we additionally used multivariate models of analysis of the prediction of ADHD by various CBCL problem scales. It was hypothesized that the DSM- ADH-scale is superior to the original CBCL attention problem scale and to a multivariate model including different empirical CBCL scales.

4.3 Methods

4.3.1 Participants

A community based sample of 392 subjects (217 boys, 175 girls) aged 6 to 17 years (mean = 12.6, SD = 2.59 years) and an outpatient sample of the same size matched for sex and age (mean = 12.6, SD = 2.64 years) was examined. Thus, the total sample contained 784 children and adolescents for whom parent-rated CBCL data and clinical diagnoses were available.

The community based sample data was taken from the Zurich Epidemiological Study of Child and Adolescent Psychopathology (ZESCAP) (Steinhausen, Winkler Metzke, Meier, & Kannenberg, 1998). A total number of 1964 students aged 6-17 years, living in the Canton of Zurich (Switzerland) and attending the first to the ninth grade in various types of schools were involved in the study. The cohort was a stratified randomized sample representing the 12 counties of the canton, the school grades and the types of school. A full description of the sampling procedures and characteristics has been given in a previous publication (Steinhausen, Winkler Metzke, Meier & Kannenberg, 1998). At stage one, the application of various screens including several CBCL syndrome scales allowed the differentiation between

screen-positive and screen-negative subjects for stage two of the assessment process that used structured interviews in order to arrive at clinical diagnoses.

A total of 557 students who were screen-positive and a randomized control sample of 122 screen-negative students were identified for further parental diagnostic interviews. Following mailed invitation, 416 parents were willing to co-operate. Due to missing items the final community based sample with both screening and interview assessment consisted of 392 subjects and included 319 screen-positives and 73 screen-negatives subjects. Of the 392 subjects in the sample, 111 subjects were screen-positive and 281 were screen-negative for attention problems based on the 90th percentile of the corresponding original CBCL scale. All diagnostic interviews were performed blindly to the results of the initial screening procedure. Attrition analyses showed that the 392 participating subjects did not differ significantly from the 289 drop-outs in terms of age (Mean = 11.81 vs.11.97, t = .737, df = 679, p = n.s), gender distribution (56.4 % vs. 55.4 %, Chi2 = .074 df = 1, p = n.s.) and CBCL total problem score (Mean = 28.91 vs. 30.97, t = 1.396 df = 679, p = n.s.).

A total of N = 9532 referrals to the child and adolescent psychiatry service of the canton of Zurich (Switzerland) between late 2001 and March 2006 were eligible for inclusion into the outpatient sample of the present study. Out of this cohort a random sub-sample matched for sex and age to the community sample of 392 subjects was drawn.

In order to exclude sampling effects and to improve the reliability of the results, the community and the outpatient sample were split into two subgroups each, namely, prediction and cross-validation sub-samples. Group assignment was done by random sampling controlling for age and sex distribution. Because two subjects had incomplete diagnostic information on ADHD both community based sub-samples had 196 subjects and both outpatient sub-samples had 197 subjects.

4.3.2 Measures

CBCL

The Swiss adoption (Steinhausen, Winkler Metzke, & Kannenberg, 1996) of the CBCL 4-18 1991 Profile (Achenbach, 1991) was used in both samples of the present study. The CBCL has three levels of scoring: (1) eight primary scales named withdrawn, somatic, anxious/depressed, social problems, thought problems, attention problems, delinquent, and aggressive behavior; (2) two second order scales called internalizing and externalizing and (3) a total problem score. The original DSM-oriented ADH problem scale is based on the 2001 revised version of the CBCL (Achenbach & Rescorla, 2001). The scale includes seven items but only five are consistent with the 1991 Profile. Thus, in the present study a reduced 5-item DSM attention problem scale had to be used that included the following items: (8) Can't concentrate, (10) Can't sit still, (41) Impulsive, (93) Talks too much and (104) Is too loud.

Psychiatric diagnoses

Different diagnostic criteria were used in the two samples. In the community sample, the Diagnostic Interview Schedule for Children – Parent Version (DISC 2.3) (Shaffer et al., 1993) was used and DSM-III-R criteria for diagnosis have been applied. The time frame of diagnoses was the six months period preceding the interview. In the outpatient sample, consensus diagnoses were provided in each case by a postgraduate clinician and a senior child and adolescent psychiatrist according to diagnostic criteria of the ICD-10 classification system. The best estimate procedure was used, i.e., raters used all available information including history, reports from psychological and educational testing, behavioral observations, and school reports. Out of 392 a total of 312 subjects had one or more psychiatric diagnoses. In the remaining 80 cases, psychiatric problems were either not exceeding subthreshold levels or involved developmental disorders only.

4.3.3 Statistical analyses

First, the predictive diagnostic potential of the original attention problem scale alone was compared to models resulting from a multivariate approach using several scales of the CBCL in (A) the community based prediction sample and (B) the outpatient prediction sample. If none of these two prediction models showed a superior performance of the original attention problem scale, the performance of the latter scale was compared to the performance of the DSM-ADH-scale. If a superior multivariate prediction model was found, it was selected for further analyses. Cut-off analyses were performed only for the most accurate prediction model.

In a first step, the multivariate approach was based on separate univariate logistic regression analyses for each CBCL original problem scale separately in both prediction sub-samples (Bonferroni correction: $p < 0.00625$). A scale was considered a candidate for further prediction analysis if the findings were significant in one of the regression equations in either the outpatient sample or the community sample. Next, the identified scales were used as predictors in stepwise logistic regression analyses (entry level $p = 0.05$) in order to select the best predictor or combination of predictors of ADHD separately in both sub-samples. Additional receiver operating characteristics (ROC) analyses based on the probabilities of the specified regression models were performed using the area under the curve (AUC) as a measure of the diagnostic accuracy. The AUC of each model within the different sub-samples was compared and tested for significance. For comparison of different scales within the same sample, a critical z-ratio was calculated using a formula correcting for the non-independence of the scales (Hanley & McNeil, 1983).

After identifying the most effective model, the optimal cut-off score for that model was established. Efficiency (EFF) was calculated by the sum of true positives (TP) and true negatives (TN). All of these indices are dependent on the prevalence of the disorder (P) and the level of the test (Q; prevalence of a positive test result). A method for the definition of the

optimal cut-off score of a test has been introduced by Kraemer (1992) by the calibration of sensitivity and specificity to the base rates and the calculation of a corrected efficiency index. The transformation of the ROC curve by adjusting sensitivity and specificity is called the quality ROC curve (Q-ROC). A quality index of efficiency was calculated using the following formula: dQ= [EFF – PQ - (1 - P)(1 - Q)]/[1 - PQ- (1 - P)(1 - Q)].

4.4 Results

4.4.1 Description of the samples

Table 2 provides an overview of the frequencies of psychiatric disorders in the outpatient and in the community based sample. There were 47 subjects with ADHD in the community sample and 65 subjects with ADHD in the outpatient sample. Approximately 50% of ADHD subjects in both samples had at least one comorbid psychiatric disorder.

Table 2. *Frequencies of psychiatric disorders in the community and the outpatient sample*

	Community based sample		Outpatient sample	
Psychiatric disorders	Total sample (N = 390)	ADHD sub- sample (N = 47)	Total sample (N = 392)	ADHD sub-sample (N = 65)
ADHD	47 (12.0%)	47 (100%)	65 (16.6%)	65 (100%)
Anxiety and obsessive-compulsive disorders	59 (15.1%)	9 (19.1%)	42 (10.7%)	4 (6.1%)
Tic disorders	28 (7.2%)	2 (4.3%)	4 (1.0%)	1 (1.5%)
Affective disorders	7 (1.8%)	1 (2.1%)	42 (10.7%)	2 (3.1%)
ODD	14 (3.6)	9 (19.1%)	41 (10.5%)	21 (32.3%)
CD	1 (0.3%)	0 (0%)	21 (5.4%)	2 (3.1%)
Drug abuse or dependence	2 (0.5%)	0 (0%)	10 (2.6%)	1 (1.5%)
Cases with one or more (comorbid) disorders	122 (31.1%)	22 (46.8%)	178 (45.4%)	34 (52.3%)

Means and standard deviations of the CBCL scores and the results of the two (prediction vs. cross-validation sample) by two (outpatient sample vs. community sample) by two (ADHD vs. non ADHD) MANOVA are shown in Table 3. As expected by random sampling, no significant

differences were found between the prediction and the cross-validation sub-samples. However, significant ADHD effects were detected for the DSM-ADH-scale and all CBCL original problem scales expect for withdrawn, somatic complaints and thought problems. Furthermore, significant mean differences were detected for all CBCL scales between the outpatient and the community based sample and significant interaction effects of ADHD and sample were found for all scales expect for social problems, delinquent behavior, aggressive behavior and the DSM oriented ADH-scale.

Table 3. *Means and standard deviations of CBCL syndromes scores in four groups of subjects (T-scores) and results of the 2 (sample) x 2 (random sub-sample) x 2 (ADHD) MANOVA*

Sample	Community based sample (N = 390)								Outpatient sample (N = 392)								MANOVA Between-Subjects Effects			
Random sub-sample	Prediction sample (N = 195)				Cross-validation sample (N =195)				Prediction sample (N = 196)				Cross-validation sample (N =196)				Random sub-sample effect	Sample effect	ADHD effect	Interaction effect: sample x ADHD
ADHD	no ADHD (N = 173)		ADHD (N = 22)		no ADHD (N = 170)		ADHD (N = 25)		no ADHD (N = 167)		ADHD (N = 29)		no ADHD (N = 160)		ADHD (N = 36)					
	means	SD	means	SD	means	SD	means	SD	means	SD	means	SD	means	SD	means	SD	F	F	F	F
Age	12.76	2.60	11.32	2.38	12.56	2.64	12.28	2.32	12.70	2.61	11.34	2.50	12.76	2.58	12.11	2.91				
Syndrome scales																				
Withdrawn	52.99	9.89	57.00	6.99	53.72	9.81	56.06	7.69	62.54	8.53	60.84	10.58	61.99	10.01	57.07	9.51	1.33	33.21***	0.01	10.86**
Somatic Complaints	51.80	9.65	52.37	9.58	53.05	9.14	57.74	12.34	59.17	11.51	53.92	12.77	56.96	11.84	54.47	9.24	1.27	4.70*	0.32	8.71**
Anxious / Depressed	54.91	10.39	59.08	5.37	54.37	9.70	60.35	8.25	62.26	9.84	62.26	10.80	61.67	9.74	61.46	8.63	0.03	21.96***	6.06*	6.56*
Social Problems	53.23	8.63	60.56	7.82	54.03	9.14	59.85	11.96	58.75	8.78	63.26	9.02	59.52	9.87	62.80	9.70	0.01	19.10***	30.23***	1.98
Thought Problems	53.34	8.69	57.20	9.46	52.87	8.12	59.05	9.72	59.19	9.56	58.56	12.73	59.82	10.20	57.35	9.32	0.04	10.37**	3.22	11.53**
Attention Problems	54.32	9.57	65.51	7.76	54.56	9.54	64.90	6.48	59.99	9.38	65.90	8.85	61.02	9.16	66.06	7.06	0.05	12.95***	72.90***	7.74**
Delinquent Behavior	53.19	9.23	60.85	10.32	52.77	9.18	61.67	7.66	59.27	11.48	64.88	11.34	59.09	11.52	64.26	12.08	0.00	19.37***	40.08***	1.79
Aggressive Behavior	53.59	9.99	63.88	8.35	53.85	9.85	63.97	7.34	59.24	10.39	67.23	9.93	59.41	11.11	68.28	9.84	0.14	20.09***	78.29***	0.71
DSM-oriented scale																				
ADH Problems (raw scores)	2.20	1.92	5.41	1.71	2.13	1.96	5.48	1.78	2.81	2.19	5.59	2.60	2.96	2.37	5.86	2.57	0.23	5.16*	191.53***	1.02

Note. 2 x 2 x 2 MANOVA: Wilks' Lambda = .019, F = 4480.87, df =98, p < 0.001, random sub-sample effect: Wilks' Lambda = .994, F = 0.55, df = 9, p = n.s., sample effect: Wilks' Lambda = .944, F = 5.08, df = 9, p < 0.001, ADHD effect: Wilks' Lambda = .789, F = 22.79, df = 9, p < 0.001. Interaction sample x random sub-sample effect: Wilks' Lambda = .993, F = 0.60, df = 9, p = n.s., interaction random sub-sample x ADHD effect: Wilks' Lambda = .990, F = 0.87, df = 9, p = n.s., interaction ADHD x sample effect: Wilks' Lambda = .972, F = 2.47, df = 9, p < 0.01. *** = p < 0.001, * = p < 0.01, * = p < 0.05. All CBCL scores except DSM-oriented ADH problems scores are T-Scores.

4.4.2 Logistic regression analyses

In the community based prediction sub-sample, four problem scales significantly predicted the presence of ADHD: social problems, attention problems, delinquent behavior, and aggressive behavior. The same scales also significantly predicted the presence of ADHD in the outpatient prediction sub-sample with the exception of the delinquent behavior scale which failed to be significant. Stepwise multivariate analyses in the community prediction sample resulted in a prediction model including the aggression and attention problem scale. The same type of analysis lead to a model that was based only on the aggressive behavior scale as single predictive variable in the outpatient prediction sample. Unstandardized regression coefficients, standard errors of the unstandardized regression coefficients and Wald T-test scores for both prediction models are displayed in Table 4.

Table 4. *Results of the stepwise logistic regression analyses in the prediction of ADHD*

Predictors	*4.4.2.1*	SE	WALD T	Df	Sig.
Model for the prediction community sub-sample					
Constant	-5.25	.84	43.27	1	.00
Attention Problems	.32	.09	13.74	1	.00
Aggressive Behavior	.14	.05	9.54	1	.00
Model for the prediction outpatient sub-sample					
Constant	-3.39	.49	48.44	1	.00
Aggressive Behavior	.12	.03	18.43	1	.00

Note. B = unstandardized regression coefficient.

4.4.3 ROC Analyses

ROC graphs for the prediction of ADHD in the community based prediction sub-sample and outpatient sub-sample are displayed in figures 1 and 2. The original attention problem scale, the aggression scale (resulting from multivariate logistic regression in the outpatient prediction sample), and the multivariate model including the aggression and the original

attention problem scale (resulting from multivariate logistic regression in the community based prediction sample) were compared by use of ROC analyses. In the community prediction sub-sample, the model including the original attention problem and aggressive behavior scales showed the highest AUC (.880) whereas in the outpatient prediction sub-sample the model including the aggressive behavior scale as a single variable showed the highest AUC (.734).

Table 5. *Comparison of the Area under the Curve (AUC) for the original attention problem scale and the two multivariate prediction models in four sub-samples*

Model (AUC)	*z*	Sig.
Community prediction sub-sample		
ATT (.839) vs. M2 AGG (.817)	0.35	n.s.
M2 AGG (.817) vs. M1 AGG + ATT (.880)	1.48	n.s.
M1 AGG + ATT (.880) vs. ATT (.839)	1.23	.n.s.
Community cross-validation sub-sample		
ATT (.829) vs. M2 AGG (.806)	0.50	n.s.
M2 AGG (.806) vs. M1 AGG + ATT (.853)	1.29	n.s.
M1 AGG + ATT (.853) vs. ATT (.829)	0.84	.n.s.
Outpatient prediction sub-sample		
ATT (.682) vs. M2 AGG (.734)	0.93	n.s.
M2 AGG (.734) vs. M1 AGG + ATT (.722)	0.28	n.s.
M1 AGG + ATT (.722) vs. ATT (.682)	1.05	.n.s.
Community cross-validation sub-sample		
ATT (.691) vs. M2 AGG (.743)	1.04	n.s.
M2 AGG (.743) vs. M1 AGG + ATT (.738)	0.14	n.s.
M1 AGG + ATT (.738) vs. ATT (.691)	1.39	.n.s.

Note. AUC = Area under the curve, ATT = Original Attention problem scale, M2 AGG = Logistic regression model based on the outpatient prediction sample including the aggressive behavior scale, M1 AGG + ATT = Logistic regression model based on the community sample including the attention problem and the aggressive behavior scale.

However, no significant differences were found between the AUC based on the original attention problem scale and on the two multivariate prediction models in both sub-samples. These results were confirmed in the corresponding cross-validation sub-samples. Table 5 shows the comparison of the AUC in all four sub-samples.

On the other hand, significant differences between findings in the community and the outpatient sample were detected. The original attention problem scale led to a significantly better prediction of ADHD in the community based prediction sub-sample than in the outpatient prediction sub-sample ($z = 2.32$, $p < 0.05$). This finding was confirmed in the cross-validation sub-samples ($z = 2.38$ $p < 0.05$). A significantly larger AUC was also found for the multivariate model in the community based sample, which included the aggressive behavior and attention problem scale (prediction sub-samples $z = 2.55$, $p < 0.05$, cross-validation sub-samples $z = 2.05$, $p < 0.05$). However, this was not true for the aggressive behavior scale alone (prediction samples $z = 1.18$, p = n.s., cross-validation sample $z = 1.04$, p = n.s.).

In a second step, the original attention problem scale was compared with the 5-item DSM ADH scale. The 5-item DSM-ADH-scale showed the highest AUC in both samples (Figures 2 and 3) which was significantly higher than the AUC of the original attention problem scale in the outpatient sample (prediction sub-sample $z = 3.07$, $p < 0.05$, cross-validation sub-sample $z = 3.25$, $p < 0.05$) but not in the community sample (prediction sub-sample $z = 1.51$, p = n.s., cross-validation sub-sample $z = 1.955$, p = n.s.). The AUC of the DSM-ADH-scale in the community and in the outpatient sample did not differ significantly from each other (prediction sub-sample $z = 1.59$, p = n.s., cross-validation sub-sample $z = 1.87$, p = n.s.).

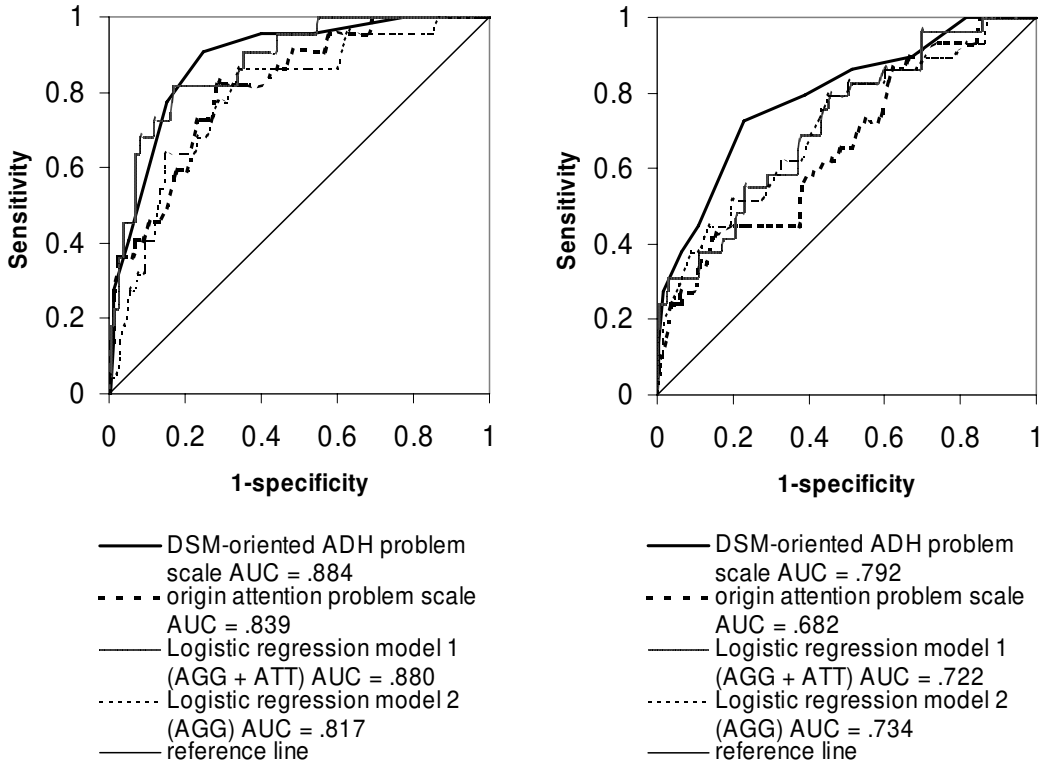

Figure 2 and 3 ROC curves for the 2 scales and 2 logistic regression models predicting ADHD in the community based prediction sample (left side) and in the outpatient prediction sample (right side)

4.4.4 Cut-off score analyses

Cut-off analyses were computed only for the DSM-ADH-scale using a quality efficiency indicator (dQ). Table 6 presents the results of the cut-off analyses of the four samples. A raw score of 5 to 6 was found to be the optimal cut-point in the outpatient and in the community sample.

Table 6. *Cut-off points analyses for the DSM-ADH-scale in four sub-samples of subjects according to a quality index of efficiency (d_Q)*

	Optimal cut-point (raw score)	Base rates	SE	SP	PPP	NPP	d_Q
Community based prediction sub-sample	5	0.22	0.77	0.85	0.40	0.97	.44
Community based cross-validation sub-sample	6	0.11	0.48	0.95	0.57	0.93	.46
Outpatient prediction sub-sample	5	0.30	0.72	0.77	0.36	0.94	.35
Outpatient cross-validation sub-sample	6	0.22	0.56	0.85	0.45	0.89	.37

Note. SP = specificity; SE = sensitivity; PPP = positive predictive power; NPP = negative predictive power; d_Q = quality index for efficiency.

4.5 Discussion

This study attempted to test the DSM-ADH-scale of the CBCL for the prediction of ADHD in comparison to the original CBCL problem scales in two different samples of referred and non-referred subjects. In both samples, the DSM-oriented scale was superior in predicting ADHD. However, in contrast to the original attention problem scale only the improvement in the outpatient sample was significant. AUCs as a measure of excellence for predicting diagnosis should be interpreted as follows: poor (50-.70) ; moderate to fair (.70-.80); good (.80-.90), and excellent (.90-1.00) (Ferdinand, 2008). Thus, the reduced 5-item DSM ADH scale showed a good prediction of ADHD in the community sample with an AUC of .88 and .89 and still showed a fair to good prediction of ADHD in the outpatient sample with an AUC of .79 and .80, respectively. Despite the reduced number of items of the present scale, these results based on ROC analyses confirm previous findings that the DSM-ADH-scale based on 7 items is an adequate instrument for diagnosing ADHD as recommended by Achenbach and Rescorla (2003; 2001).

4.5.1 Limitations of the original CBCL problem scales in predicting ADHD

The original attention problem scale and the two prediction models resulting from logistic regression analyses predicted ADHD adequately in the community based sample but not in the outpatient sample. In both samples, no superior multivariate model was detected when compared to the original attention problem scale. When looking at the diagnostic accuracy in of the original attention problem scale in the community based sample, the results from previous studies based on parental diagnostic interviews were confirmed. The AUC of .839 in the community prediction sub-sample and the AUC of .829 in the corresponding cross-validation sub-sample indicate an acceptable quality of ADHD prediction which is comparable to previous findings (Chen et al., 1994; Hudziak et al., 2004). Both of these studies included subjects from referred and non-referred recruitment sources. Thus, the present study improved the validity of the original attention problem scale for predicting ADHD in a community based sample without subjects from psychiatric institutions.

However, the present results could not confirm the diagnostic validity of the original attention problem scale for predicting ADHD in an outpatient sample. In the same way, also a multivariate model based on several CBCL scales was not found to be more accurate in predicting ADHD. Furthermore, the aggressive behavior scale was found to be most strongly related to the diagnoses of ADHD in the outpatient sample. This result may be due to the higher number of comorbid ODD disorders in the outpatient sample. It may be assumed that the presence of aggressive and oppositional behavior may have had an effect on the parents when they completed the CBCL. However, the prediction of ADHD by the aggressive behavior scale was still moderate and insufficient for clinical practice.

4.5.2 Further support of the DSM-ADH-scale

The present findings strongly imply to use the DSM-ADH-scale rather than the original attention problem scale for predicting ADHD. This recommendation is further supported by the following considerations.

First, the validity of the scale is supported by the fact that diagnoses were established in a clinical sample by the best estimate procedure with consensus diagnoses and the inclusion of different sources of information (children, parents, teachers etc.) Thus, the DSM-ADH-scale can be recommended for use in clinical settings with a comprehensive and multimodal diagnostic assessment approach.

Secondly, the results support the validity of the DSM-ADH-scale even for predicting diagnoses based on ICD-10 criteria of Hyperkinetic Disorder (HD) which is the equivalent to the DSM-based term of ADHD. So far, no study before has tested the prediction of these ICD-10 HD criteria by use of the CBCL.

Thirdly, the prediction of ADHD by the DSM-ADH-scale is accurate although the non-ADHD subjects in both tested samples are strongly affected by other psychiatric disorders. In the present outpatient sample, 45% of the subjects had at least one psychiatric disorder. When including further ICD-10 psychiatric disorders, approximately 80% of the subjects in the outpatient sample have at least one mental disorder. In this sample, the DSM-ADH-scale was able to discriminate ADHD from other psychiatric disorders that include attention problems also very frequently. Thus, it may be concluded that the reduced list of 5 items in the present DSM-ADH-scale is more sensitive for the identification of ADHD than the original attention problem scale of the CBCL which includes further items that correspond more strongly to other psychiatric disorders.

4.5.3 Cut-off points and recommendations for clinical use

Cut-off point analyses indicate a raw score of 5 to 6 on the 5-item DSM-ADH-scale as the optimal cut-off point weighting the impact of identifying cases and non-cases as equivalent. Due to the low base rates of the disorder in both samples the positive predictive power of the original attention problem scale was lower than in previous studies (Chen et al., 1994; Hudziak et al., 2004). Additionally, these results may be due to methodological differences (Gray, 2004) because these studies used different recruitment sources of healthy controls and ADHD subjects. The so-called spectrum bias (Knotterus, 2002) may have led to higher

sensitivity values. By assuming higher costs of false negative classifications due to the serious handicaps of non-treated ADHD (Barkley, Fischer, Smallish, & Fletcher, 2004) a lower cut-off score with better sensitivities should be considered. Therefore a cut-off point of 5 can be recommended as an initial starting point in clinical diagnostic assessment on ADHD. Additionally, further information on age of onset, continuity, impairment, specificity of symptoms and rater agreement should be considered in order to arrive at the final diagnosis.

4.5.4 Limitations

The present study has some limitations. First, different criteria of ADHD diagnosis and assessment procedures in the community and the clinical sample limit the comparability of ROC analyses results. Therefore, the diverging results can not be strictly attributed only to the fact that different samples of referred and non-referred subjects had been used in the present study. Secondly, the present results based on the community sample may not generalize to other populations because subjects recruited for the present study were more strongly affected by various emotional and behavioral problems due to the multilevel screening process in the basic epidemiological study. Furthermore, no information on DSM-IV criteria of inattentive or hyperactive/impulsive subtypes was available. Because of the strong convergence of DSM-III-R and DSM-IV criteria of ADHD combined type it can be assumed that the present results would have been similar if DSM-IV criteria of diagnoses would have been used. Thirdly, no formal information on reliability of clinical ICD-10 diagnoses was available. In clinical settings, interviewers typically follow their initial diagnostic hypothesis by asking increasingly more specific questions in order to rule in or to rule out certain diagnoses (Doss, 2005). In the present study, diagnoses by postgraduate clinicians were based on clinical interviews with parents and children and included also teacher information. In each case, these diagnoses were confirmed by senior clinical experts so that the best estimate procedure was performed. In addition, clinicians were also not blind to CBCL findings including information on attention problems.

4.6 Conclusion

A 5-item DSM-oriented ADH-problem scale based on the 1991 CBCL profile can be recommended for screening of ADHD both in community and clinical samples. In contrast, the original attention problem scale was not suited for the identification of ADHD in an outpatient sample referred for various psychiatric disorders.

4.7 References

Achenbach, T. M. (1991). Manual for the Child Behavior Check List/4-18 and 1991 Profile. Burlington, VT: Department of Psychiatry, University of Vermont.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). DSM-oriented and empirically based approaches to constructing scales from the same item pools. Journal of Clinical Child and Adolescent Psychology, 32(3), 328-340.

Achenbach, T. M., & Rescorla, L. A. (2001). Manual for the School-Age Forms and Profiles. Child Behavior Checklist. Teacher's Report Form. Youth Self-Report. An Integrated System of Multi-informant Assessment. Burlington: Library of Congress.

Barkley, R. A., Fischer, M., Smallish, L., & Fletcher, K. (2004). Young adult follow-up of hyperactive children: antisocial activities and drug use. Journal of Child Psychology and Psychiatry, 45, 195-211.

Biederman, J., Faraone, S. V., Doyle, A., Lehman, B. K., Kraus, I., Perrin, J., et al. (1993). Convergence of the Child Behavior Checklist with structured interview-based psychiatric diagnoses of ADHD children with and without comorbidity. Journal of Child Psychology and Psychiatry, 34(7), 1241-1251.

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist scales for attention-deficit hyperactivity disorder: a receiver-operating characteristic analysis. Journal of Consulting and Clinical Psychology, 62, 1017-1025.

Doss, A. J. (2005). Evidence-based diagnosis: incorporating diagnostic instruments into clinical practice. Journal of the American Academy of Child and Adolescent Psychiatry, 44, 947-952.

Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and criterion-related validity of the Behavior Assessment System for Children-Parent Rating Scale. Journal of Clinical Child Psychology, 26, 276-284.

Edelbrock, C., & Costello, A. J. (1988). Convergence between statistically derived behavior problem syndromes and child psychiatric diagnoses. Journal of Abnormal Child Psychology, 16, 219-231.

Eiraldi, R. B., Power, T. J., Karustis, J. L., & Goldstein, S. G. (2000). Assessing ADHD and comorbid disorders in children: the Child Behavior Checklist and the Devereux Scales of Mental Disorders. Journal of Clinical Child Psychology, 29, 3-16.

Ferdinand, R. F. (2008). Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. Journal of Anxiety Disorders, 22(1), 126-134.

Ferdinand, R. F., Visser, J. H., Hoogerheide, K. N., van der Ende, J., Kasius, M. C., Koot, H. M., et al. (2004). Improving estimation of the prognosis of childhood psychopathology; combination of DSM-III-R/DISC diagnoses and CBCL scores. Journal of Child Psychology and Psychiatry, 45, 599-608.

Gray, G. E. (2004). Evidence Based Psychiatry. Washington DC: American Psychiatric Publishing.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148, 839-843.

Hudziak, J. J., Copeland, W., Stanger, C., & Wadsworth, M. (2004). Screening for DSM-IV externalizing disorders with the Child Behavior Checklist: a receiver-operating characteristic analysis. Journal of Child Psychology and Psychiatry, 45, 1299-1307.

Jensen, P. S., Salzberg, A. D., Richters, J. E., & Watanabe, H. K. (1993). Scales, diagnoses, and child psychopathology: I. CBCL and DISC relationships. Journal of the American Academy of Child and Adolescent Psychiatry, 32, 397-406.

Kazdin, A. E., & Heidish, I. E. (1984). Convergence of clinically derived diagnoses and parent checklists among inpatient children. Journal of Abnormal Child Psychology, 12, 421-435.

Knotterus, J. A. (2002). The Evidence Base of Clinical Diagnosis. London: BMJ Books.

Kraemer, H. C. (1992). Evaluating medical tests. Objective and quantitative guidelines. Newbury Park: Sage Publications, Inc.

Kraemer, H. C., Noda, A., & O'Hara, R. (2004). Categorical versus dimensional approaches to diagnosis: Methodological challenges. Journal of Psychiatric Research, 38, 17-25.

Lampert, T. L., Polanczyk, G., Tramontina, S., Mardini, V., & Rohde, L. A. (2004). Diagnostic performance of the CBCL-Attention Problem Scale as a screening measure in a sample of Brazilian children with ADHD. J Atten Disord, 8(2), 63-71.

Leon, A. C., Olfson, M., Weissman, M. M., Portera, L., & Sheehan, D. V. (1996). Evaluation of screens for mental disorders in primary care: methodological issues. Psychopharmacological Bulletin, 32(3), 353-361.

Ostrander, R., Weinfurt, K. P., Yarnold, P. R., & August, G. J. (1998). Diagnosing attention deficit disorders with the Behavioral Assessment System for Children and the Child Behavior Checklist: test and construct validity analyses using optimal discriminant classification trees. Journal of Consulting and Clinical Psychology, 66, 660-672.

Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Piacentini, J., Davies, M., et al. (1993). The Diagnostic Interview Schedule for Children-Revised Version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability. Journal of the American Academy of Child and Adolescent Psychiatry, 32, 643-650.

Steingard, R., Biederman, J., Doyle, A., & Sprich-Buckminster, S. (1992). Psychiatric comorbidity in attention deficit disorder: impact on the interpretation of Child Behavior Checklist results. Journal of the American Academy of Child and Adolescent Psychiatry, 31(3), 449-454.

Steinhausen, H. C., Winkler Metzke, C., & Kannenberg, R. (1996). Handbuch: Elternfragebogen über das Verhalten von Kindern und Jugendlichen. Die Zürcher Ergebnisse zur deutschen Fassung der Child Behavior Checklist (CBCL). Zürich: Zentrum für Kinder- und Jugendpsychiatrie der Universität Zürich.

Steinhausen, H. C., Winkler Metzke, C., Meier, M., & Kannenberg, R. (1997). Behavioral and emotional problems reported by parents for ages 6 to 17 in a Swiss epidemiological study. European Child and Adolescent Psychiatry, 6, 136-141.

Steinhausen, H. C., Winkler Metzke, C., Meier, M., & Kannenberg, R. (1998). Prevalence of child and adolescent psychiatric disorders: the Zurich Epidemiological Study. Acta Psychiatrica Scandinavica, 98, 262-271.

Zelko, F. A. (1991). Comparison of parent-completed behavior rating scales: differentiating boys with ADD from psychiatric and normal controls. Journal of Development & Behavioral Pediatrics, 12(1), 31-37.

5 Study 2: Prediction of major affective disorders in adolescents by self - report measures[2]

5.1 Abstract

Background: The Youth Self - Report (YSR) has been used widely as a screening instrument for adolescent psychopathology. The present study aimed at a test of the diagnostic accuracy of the various YSR – scales including a DSM-oriented affective problem scale (YSR AFF) in the prediction of depressive episodes and a comparison with results based on the Center of Epidemiologic Studies-Depression Scale (CES-D).

Methods: A consecutive clinical sample of 140 adolescents diagnosed with major depressive episodes according to ICD-10 criteria was compared to a sample of 140 non-referred controls matched by age and sex from a community survey. All subjects responded both to the YSR and CES-D. Diagnoses were provided by the treating clinicians. Receiver Operating Characteristics (ROC) analyses were performed and cut-off scores were calculated based on quality efficiency statistics.

Results: The YSR AFF scale was found to have high diagnostic accuracy and showed quite comparable results to the CES-D scale. None of the other multivariate model showed a better performance in the identification of major depression disorders. Based on quality efficiency indicator analyses, scores between 5 and 9 on the YSR AFF - scale and between 12 and 31 on the CES-D scale served best in the prediction of clinical depressive episodes in adolescents.

Limitations: No formal reliability test of the diagnoses was available.

Conclusion: The DSM oriented YSR AFF scale shows a high diagnostic accuracy and can be recommended for the clinical assessment of depression in adolescents.

Keywords: Adolescence, affective disorders, depression, screening, YSR

---

[2] Aebi, M., Winkler Metzke, C. & Steinhausen, H.-C. (2008) Prediction of major affective disorders in adolescents by self-report measures, *Journal of Affective Disorders*, doi:10.1016/j.jad.2008.09.017

5.2 Introduction

A review of recent epidemiological studies showed that adolescent depression is quite common with prevalence rate ranging between 1.8% and 5.9% (Costello et al., 2006; Lewinsohn et al., 1993; Roberts et al., 2000; Steinhausen et al., 1997) and further risks of abnormal psychosocial and mental functioning in young adulthood (Steinhausen et al., 2006). Diagnosis is predominantly based on detailed interviews with the adolescent. However, as with most adolescent psychopathological disorders, questionnaires add considerable information to the assessment process.

The Center of Epidemiologic Studies-Depression Scale (CES-D, Radloff, 1977) is a well-known and suitable screening instrument for affective disorders. Diagnostic accuracy of the CES-D for DSM diagnosis of major depression has been shown in community and outpatient adult (Roberts and Vernon, 1983; Weissman et al., 1977) and adolescent samples (Garrison et al., 1991; Roberts et al., 1991; Yang et al., 2004).

However, given the frequent comorbidity in adolescent depression (Angold and Costello, 1993; Kovacs et al., 1989; Kovacs et al., 1988), a multidimensional questionnaire could provide more relevant clinical information, rather, than a single scale only. The Youth - Self Report (YSR, Achenbach, 1991b) is a multidimensional questionnaire for a broad range of behavioral and emotional problems in adolescents. The YSR has been used widely both for screening of mental health problems in the community and in clinical settings (e. g. Rohde et al., 2004; Saxena et al., 2005). Usually, categorical diagnoses are needed in clinical settings both for medical and administrative purposes. In contrast, the problem syndrome scales of the YSR have been empirically defined and reflect a dimensional approach to psychopathology. Whereas some of the YSR - scales bear an a priori resemblance to certain psychiatric diagnoses, others do not.

So far, three studies have addressed the association between YSR- syndrome scales and interview based DSM III / DSM-III-R – diagnoses of depression based on the child version of the Diagnostic Interview Schedule for Children (DISC-C, Shaffer et al., 1996) in an outpatient (Gould et al., 1993; Morgan and Cauce, 1999) and in an inpatient sample (Weinstein et al., 1990). Statistically meaningful though rather weak correlations with the diagnoses of depression and dysthymia were found for the somatic complaint scale (Morgan and Cauce, 1999) and the anxious depressed scale (Gould et al., 1993; Morgan and Cauce, 1999) of the YSR. The best prediction of depressive disorders was found when both scores were at least in the borderline range (hit rate 85%, kappa = .44; Morgan and Cauce, 1999). An inpatient sample of 160 adolescents subjects with affective disorders did not show a specific YSR – profile (Weinstein et al., 1990). Next to the hypothesized anxiety/depressed scale, also various other scales showed significantly higher mean scores in depressed than in non - depressed subjects. According to the authors, these results may have been due to possible comorbid symptoms of affective disorders.

In order to overcome the insufficient prediction of psychiatric diagnoses by the YSR and the parallel parent version (Child Behavior Checklist; Achenbach, 1991a), alternative scoring systems have been proposed and recommended for the diagnostic prediction of DSM-IV diagnoses (Krol et al., 2001; Lengua et al., 2001). In addition, DSM-oriented scales have been introduced recently in the 2001 revision of the YSR (Achenbach et al., 2003; Achenbach and Rescorla, 2001). These new scales showed good psychometric properties. New cut-off-points for the DSM-scales equivalent to the corresponding empirical scales were introduced and were defined as T = 65 for borderline and T = 69 for clinical problems. The DSM-oriented problem scales were found to correlate significantly with clinical DSM-IV diagnoses (Achenbach and Rescorla, 2001). Among these DSM – oriented scales there is also an affective problem (YSR AFF) scale which is consisting of 13 items.

So far, only two recent studies have addressed the diagnostic accuracy of the recently developed DSM-oriented YSR scales in the prediction of adolescent clinical depression. One

study assessed a sample of 196 incarcerated male adolescents and used receiver operating characteristic analyses (ROC) with the area under the curve (AUC) as a measure of discriminative diagnostic potential of the scale (Vreugdenhil et al., 2006). The results of this study showed that the prediction of DSM-IV diagnoses based on the DISC-C by the DSM-oriented scales was not better than by the original scales. Clinical internalizing disorders were not detected by the corresponding YSR scales. In the prediction of affective disorders, the highest AUC was found for the DSM-oriented oppositional problem scale (AUC = .77). In contrast, the YSR AFF scale (AUC = .65) was not efficient in identifying major depression in this specific sample of incarcerated boys.

Recently, Ferdinand (2008) has used the YSR AFF scale and ROC analyses in a referred sample of 150 adolescents aged 11-18-years in order to predict interview based major depression and dysthymia based on the Anxiety Disorders Interview Schedule for Children (ADIS C/P; Silverman et al., 2001). Different results were found for diagnoses based on parent/child rated impairment and diagnoses based on clinical severity ratings. Whereas the prediction of depression (AUC = .91) and dysthymia (AUC = .87) worked well according to the adolescents' self-report or the parents rating of impairment, the prediction of diagnoses based on clinical severity ratings for depression (AUC = .76) and dysthymia (AUC = .81) was only marginally lower.

In contrast to the study by Vreugdenhil et al. (2006), the results from the latter study confirm the diagnostic accuracy of the YSR AFF scale. This discrepancy may be due to a sample effect. Further support for the validity of the YSR AFF scale comes from a study by van Lang, Ferdinand, Oldehinkel, Ormel, & Verhulst (2005) showing a high convergence of the YSR YSR AFF scale with the Revised Children's Anxiety and Depression scale (RCADS, Chorpita et al., 2000). However, despite these promising though limited findings further affirmative research is needed given the fact that the sample by Ferdinand (2008) included only very few patients with the diagnosis of depression according to child impairment rating (N = 25)

and according to clinical severity rating (N = 9). In addition, there are some shortcomings in these findings.

First, the cut-off scores proposed by Achenbach and Rescorla (2001) for the empirical and the DSM-oriented scales have not been defined by addressing specificity and sensitivity measures and ROC analyses. In contrast, ROC analyses were performed to define the cut-off scores of the total problem, the internalizing, and the externalizing scales. Furthermore, the borderline range and the clinical range of the DSM-oriented problem scales are reflecting the empirical distribution of the entire representation and validation sample and are defined arbitrarily by T-scores of 65 and 69, respectively. However, these cut-off-points do not necessarily allow the prediction of specific diagnostic categories. Because prevalence rates of disorders vary, it seems questionable to define a common cut-off range of scales for different diagnostic constructs. On the contrary, specific YSR - scales cut-off scores for the prediction of depression may be necessary. Secondly, most of the preceding studies did not compare the diagnostic accuracy of the YSR scales with more specific symptom - scales with well - proven diagnostic utility in different samples.

The present study aimed at testing the diagnostic accuracy of both the empirical and the DSM-oriented scales of the YSR in the identification of adolescents from a clinical sample who had been diagnosed with depressive episodes according to ICD-10 criteria compared to a sample of non-referred adolescents from a community survey. It was hypothesized that the YSR AFF scale would have the best predictive power in the identification of subjects with depressive disorders compared to (a) the original anxious/depressed scale, (b) to a multivariate model including various empirical YSR-scales and (c) to the remaining DSM-oriented scales. In addition, it was expected that the findings based on the YSR AFF scale would be comparable to those based on the CES-D which is a suitable diagnostic tool for the identification of depression.

5.3 Methods

5.3.1 Participants

Between 2005 and 2006 a total of 5791 children and adolescents were admitted to the Child and Adolescent Psychiatric Service (CAPS) of the Canton of Zurich. A total of 260 (4.5%) of these subjects fulfilled criteria of ICD-10 diagnosis of depressive episode (F32.0, F32.1, F32.2) or recidivism depressive episode (F33.0, F33.1, F33.2). Consensus diagnoses were provided in each case by a postgraduate clinician and a senior child and adolescent psychiatrist. The best evidence practice model was applied in terms of using all available information including history, reports from psychological and educational testing, behavioral observations, clinical questionnaires, and school reports. Clinicians were blind to the scores of the YSR DSM-oriented scales.

Out of this clinical sub-sample, only those subjects were included who fulfilled clinical criteria of a depressive episode and who also responded to the CES-D and the YSR in the initial diagnostic assessment. 12 subjects did not fulfill age criteria (< 11 years) and therefore did not respond to the CES-D and YSR. In 98 cases no complete diagnostic assessment was made due to emergency outplacement into another institution or due to repeated admissions to our clinic. If a patient had multiple assessment episodes between 2005 and 2006 only the first entry was considered. 10 cases had more than 10% percent missing items in either the YSR or the CES-D and were excluded from the present sample. The final clinical sub-sample consisted of 140 subjects with depressed episodes (46 males, 94 females) with a mean age of 15.5 years. Attrition analyses showed that the 140 participating subjects did not differ significantly from the 108 drop-outs in terms of age (Mean = 15.40 vs.15.03, t = -1.613, df = 246, p = n.s) and gender distribution (32.8 % vs. 43.5 % males, Chi2 = .086, df = 1, p = n.s.).

The non-referred sub-sample was taken from the Zurich Epidemiological Study of Child and Adolescent Psychopathology (ZESCAP, Steinhausen et al., 1998). A total number of N=1964 students aged 6-17 years, living in the Canton of Zurich (Switzerland), attending the first to

the ninth grade in various types of schools were involved in the study. The 11 – 17 year olds (N=1110) responded to self - report measures including the YSR and CES-D. Out of this community sample, a random sub-sample matched for sex and age to the clinical sub-sample of 140 subjects was drawn.

5.3.2 Measures

Youth Self Report (YSR)

The Youth Self Report (YSR, Achenbach, 1991b; Achenbach and Rescorla, 2001) is consisting of 118 items leading to a total problem score, two second order scales (internalizing and externalizing) and eight empirically derived first order scales addressing a broad spectrum of emotional and behavioral symptoms which have been present in the past six months. The eight primary scales are labeled withdrawn, somatic complaints, anxious/depressed, social problems, thought problems, attention problems, delinquent behaviour, and aggressive behaviour. Cut-off scores of T = 67 for the borderline range and T = 70 for the clinical range have been proposed for the eight primary scales by the authors. In the 2001 revision of the YSR, six additional DSM-oriented scales have been introduced and lower cut-off scores for the borderline (T = 65) and the clinical range (T = 69) on both the empirical and DSM-oriented scales have been recommended (Achenbach and Rescorla, 2001). Reliability and validity have been shown to be good both for the original US version (Achenbach, 1991b; Achenbach and Rescorla, 2001) and the Swiss version (Steinhausen et al., 1996) of the YSR. The latter has been used in the current study. No T-scores were available for the recently developed DSM-oriented scales so that all statistics are based on raw scores.

Center of Epidemiologic Studies-Depression Scale (CES-D)

The Center of Epidemiologic Studies-Depression Scale (CES-D) is a self rating measure of depressive symptoms occurring during the last week. The scale consists of 20 items with a four point rating scale ranging from zero (rarely, less than 1 day) to three (most time, 5-7

days). In contrast to the four - factor structure that was found originally by Radloff (1977), the Swiss adaptation (Steinhausen and Winkler Metzke, 2000) of the German version (Hautzinger and Bailer, 1993) found strong evidence for a general factor including all 20 items. Thus, in the present study the total score of the CES-D was used.

5.3.3 Data analyses

In order to study the diagnostic accuracy in the prediction of depressive episodes, ROC analyses were performed separately for each empirical and DSM-oriented scale. Multivariate prediction models of depressive episodes used only the empirical YSR scales because the other DSM-oriented YSR - scales focus on other diagnostic constructs than depression. Stepwise multivariate logistic regression analyses were performed including all empirical scales which in a previous step were significantly related to the diagnosis of depressive episodes in separate univariate logistic regression models (including Bonferroni correction: $p < 0.0031$). In the present study all scales except the aggressive behavior scale were included. The calculated probabilities resulting from the final multivariate logistic regression model were, thereafter, submitted to additional ROC analyses. Thus, it was possible to compare the various results based on multivariate analyses with the findings based on single YSR - scales by using the Area under the Curve (AUC) as an overall indicator of discriminative power between subjects with depressive episodes and non-referred subjects. For comparison of different scales within the same sample, a critical z-ratio was calculated using a formula correcting for the non-independence of the scales (Hanley and McNeil, 1983).

For methodological reasons no random splitting in a prediction sub-sample and a cross-validation sub-sample was considered in the present study. A potential sampling bias would have happened before and would have affected both sub-samples equally. The sampling error would be increased by the reduced number of subjects in the sub-samples.

After identifying the most effective model, the optimal cut-off score for the respective model was established. Efficiency (EFF) was calculated by the sum of true positives (TP) and true negatives (TN). In order to correct EFF for independence of the base rate (P) in the sample and to take into account the rate of a positive test result (Q), a quality index of efficiency was calculated using the following formula: $dQ = [EFF – PQ - (1 - P)(1 - Q)]/[1 - PQ- (1 - P)(1 - Q)]$ (Kraemer, 1992).

5.4 Results

Among the 140 adolescents with depressive episodes, 83 (59%) had no comorbid disorder. The remaining 57 subjects showed the following comorbid diagnoses: 10 (18%) had attention-deficit-hyperactivity disorders, 3 (5%) oppositional defiant disorders, 2 (4%) conduct disorders, 11 (19%) anxiety disorders, 2 (4%) obsessive-compulsive disorders, 14 (25%) eating disorders, 8 (14%) substance use disorders and 20 (35%) another psychiatric ICD-10 disorder. Of these 57 subjects, 42 (74%) had one diagnosis and 15 (26%) had two ore more diagnoses.

Means and standard deviations of the CES-D score, the empirical and the DSM-oriented YSR scores are shown in Table 7. Internal consistency as measured by Cronbach's alpha was .85 of the YSR AFF scale and .83 for the CES-D scale, respectively. The scores of the two scales were strongly correlated ($r = .80$, $p < 0.001$).

Table 7. *Means and standard deviations (raw scores) of empirical and DSM-oriented YSR scores in two groups of subjects*

	Sample (N = 280)			
Diagnosis	no depression (N = 140)		depression (N = 140)	
	means	SD	means	SD
Empirical YSR scales				
Withdrawn	2.49	2.51	6.79	3.05
Somatic Complaints	2.75	2.49	5.36	3.14
Anxious / Depressed	4.97	4.27	13.82	6.61
Social Problems	1.76	2.31	3.88	2.94
Thought Problems	2.06	1.89	4.45	2.82
Attention Problems	3.94	2.67	6.83	3.06
Delinquent Behavior	3.03	2.25	5.57	3.32
Aggressive Behavior	7.35	4.01	9.42	5.22
DSM-oriented YSR scales				
Affective Problems	3.32	3.11	10.71	4.54
Anxiety Problems	2.23	2.00	5.09	2.67
Somatic Problems	1.98	1.97	3.75	2.48
ADH Problems	2.55	1.64	3.50	1.84
OD Problems	2.46	1.77	3.55	2.27
Conduct Problems	2.85	2.17	5.13	3.64
Reference scale				
CES-D	11.29	8.09	34.19	11.60

Note. All scores are raw values.

Table 8 show the results of the ROC analyses of all YSR scales, the CES-D scale and the final multivariate prediction model based on stepwise logistic regression analysis. The latter model consists of three scales, namely, the anxious/depressed, the withdrawn and the delinquent behavior scales (see Table 9). Pseudo r-square statistics indicated that approximately 60% of the variation can be explained by the model (Nagelkerke R^2 = 0.580).

A total of 81.8% of the subjects were predicted correctly using a cut-value of .5 of the prediction model.

Table 8. *Results based on ROC analyses with area under the curve (AUC) of the empirical and the DSM-oriented YSR – scales, the CES-D scale and a multivariate prediction model in the prediction of ICD-10 depressive episodes*

Sample (n = 280)	AUC	95% CI	Sign. Deviation from Index Scale
Index scale			
DSM oriented affective Problems	.907	0.872-0.942	--
Empirical YSR scales			
Withdrawn	.862	0.819-0.905	$p < 0.05$
Somatic Complaints	.744	0.687-0.802	$p < 0.05$
Anxious/Depressed	.866	0.823-0.909	$p < 0.05$
Social Problems	.746	0.689-0.804	$p < 0.05$
Thought Problems	.759	0.701-0.817	$p < 0.05$
Attention Problems	.766	0.710-0.821	$p < 0.05$
Delinquent Behavior	.735	0.675-0.794	$p < 0.05$
Aggressive Behavior	.614	0.548-0.679	$p < 0.05$
Other DSM-oriented YSR scales			
Anxiety Problems	.803	0.752-0.854	$p < 0.05$
Somatic Problems	.710	0.649-0.771	$p < 0.05$
Attention-Deficit-Hyperactivity Problems	.626	0.561-0.691	$p < 0.05$
Oppositional Defiant Problems	.638	0.573-0.702	$p < 0.05$
Conduct Problems	.702	0.641-0.764	$p < 0.05$
Multivariate prediction model			
Withdrawn x Anxious/Depressed x Delinquent Beh.	.902	0.866-0.937	n.s.
Reference scale			
CES-D	.939	0.913-0.964	$p < 0.05$

Note. All scales showed significant deviance of AUC from random prediction (AUC = .5).

The ROC – analyses showed that besides the CES-D scale (AUC = .939) the YSR AFF-scale had the highest AUC values (AUC = .907). The YSR AFF - scale was superior to all remaining DSM-oriented problem scales and to all empirical YSR syndrome scales. No significant AUC differences were found between the YSR AFF - scale and the multivariate prediction model. The AUC of the YSR AFF scale was marginally smaller than the AUC of the CES-D scale ($z = -1.971$, $p < 0.05$).

Table 9. *Prediction of depressive episodes by YSR empirical scales based on stepwise logistic regression analysis*

YSR – scale	*5.4.1.1*	SE	WALD	df	Sig.	OR
Withdrawn	0.318	0.078	16.645	1	.000	1.374
Anxious/Depressed	0.135	0.039	11.651	1	.001	1.144
Delinquent Behavior	0.205	0.064	10.395	1	.001	1.228

Note. *B* = unstandardized regression coefficent.

Additionally sex and age differences in the prediction of depressive episodes were analyzed. Adolescent females showed higher AUC (YSR AFF AUC = .928, CES-D AUC = .966) compared to adolescent males (YSR AFF AUC = .884, CES-D AUC = .888) The difference was significant for the CES-D scale ($z = -2.242$, $p < .05$) but not for the YSR AFF scale ($z = -1.105$, p = n.s.). No significant statistical differences in both scales were found between AUC of subjects younger than 16 years (YSR AFF scale AUC = .916, CES-D AUC = .941) and older than 16 years (YSR AFF scale AUC = .900, CES-D AUC = .935).

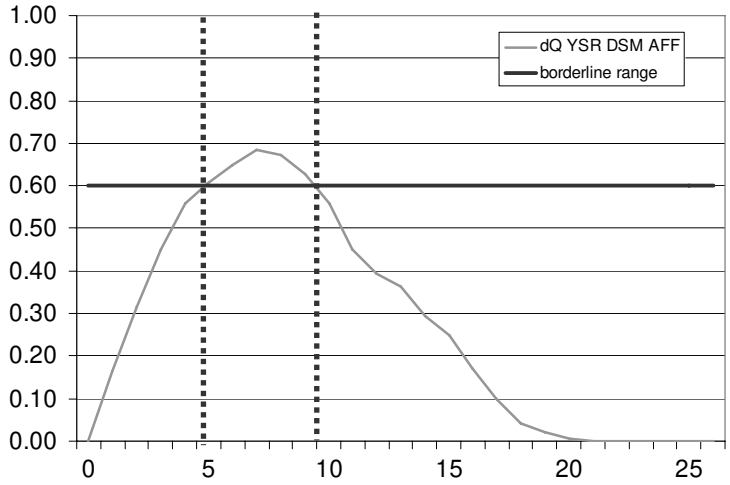

Figure 4 Quality efficiency indicator (d_Q) for all raw values of the YSR AFF – scale. The acceptable diagnostic range (5-9) as indicated by $d_Q > .60$ is marked by the vertical dotted lines.

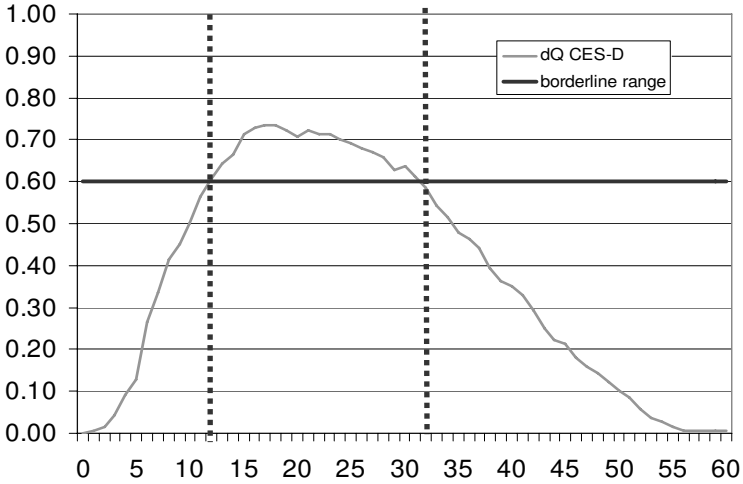

Figure 5 Quality efficiency indicator (d_Q) for all raw values of the CES-D. The acceptable diagnostic range (12-31) as indicated by $d_Q > .60$ is marked by the vertical dotted lines.

Cut-off-point analyses were performed for the YSR AFF - scale and the CES-D scale in terms of reference scale. For the YSR AFF - scale a cut-off-point of 7 was established based on efficiency statistics (dQ = .69). Overall 84% of the subjects were classified correctly by this measure. Sensitivity, specificity and positive and negative Predictive Power ranged between .81 and .87. For the CES-D, the optimal cut-off-point was 16 to 17 in the prediction

sub-sample (dQ = .74). The corresponding sensitivity and specificity figures for all specified measures were in a similar range between .81 and .93.

Table 10. *Cut-off-point analyses of the borderlines of the defined acceptable diagnostic range (d_Q > .60) and of a mid range with similar sensitivity and specificity scores for the YSR AFF scale and the CES-D scale*

	Raw score	Base rates	SE	SP	PPP	NPP	EFF	d_Q	LR +	LR-
YSR AFF scale										
Beginning range	5	0.60	0.90	0.71	0.75	0.88	0.80	0.61	3.07	0.14
End range	9	0.37	0.69	0.94	0.92	0.75	0.81	0.63	12.00	0.33
Mid range (SE ~SP)	7	0.47	0.81	0.87	0.86	0.82	0.84	0.69	6.33	0.21
CES-D										
Beginning range	12	0.66	0.96	0.64	0.73	0.95	0.80	0.61	2.7	0.06
End range	31	0.34	0.65	0.96	0.95	0.73	0.81	0.61	18.2	0.36
Mid range (SE ~SP)	21	0.50	0.86	0.86	0.86	0.86	0.86	0.72	6.05	0.16

Note. SP = specificity; SE = sensitivity; PPP = positive predictive power; NPP = negative predictive power; EFF = efficiency; d_Q = quality index for efficiency; LR+ = likelihood ratio of a positive test; LR- = likelihood ratio for a negative test.

However, in the absence of an unambiguous cut-off-point an acceptable diagnostic range defined by a quality efficiency indicator of dQ > .60 had to be tested. Landis & Koch (1977) regard reliability figures of kappa = .60 or above as an indicator of acceptable agreement. Thus, a quality efficiency indicator based on kappa = .60 or above indicates an accurate convergence of diagnosis and scale according to the defined cut-off-point. For clinical practice, different cut-off-points according to different costs and benefits for false positive and false negative diagnoses have to be considered. Therefore, it is necessary to know the range of potential acceptable cut-off-points. Figure 4 shows the quality efficiency indicator scores for all possible raw values of the YSR AFF – scale and figure 5 shows the same measures for the CES-D scale. Table 10 summarizes the results of the cut-off-point analyses for three different cut-off-points within the defined diagnostic range (borderline range and a mid range

with similar sensitivity and specificity) including the corresponding base rate, sensitivity, specificity, positive and negative predictive power and efficiency. Furthermore, the diagnostic likelihood ratios were calculated for scores above (LR+) and below (LR-) the corresponding cut-off-point. For the YSR AFF - scale a diagnostic range between 5 and 9 and for the CES-D a range between 12 and 31 was sensitive in the prediction of depressive episodes in adolescents.

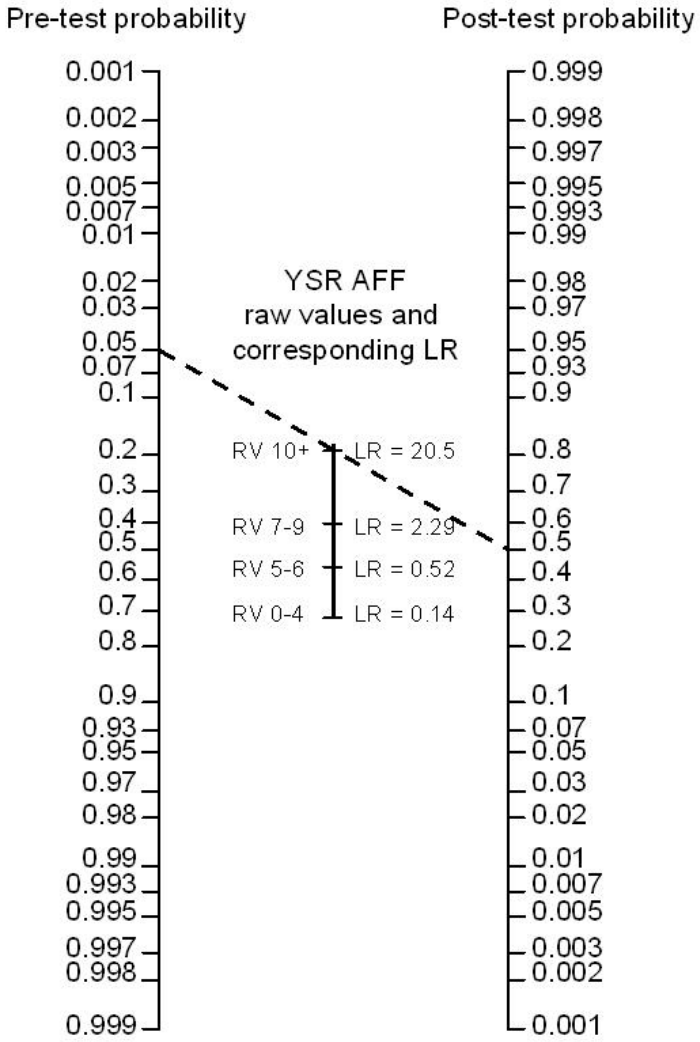

Figure 6 Nomogram for calculating post-test probabilities of ICD-10 depressive episodes by the use of the YSR DSM AFF scale.

Additionally to the proposed cut-off-points, diagnostic likelihood ratios (LR) were calculated for specific ranges of the YSR AFF scores. In clinical practice it is essential to know how a particular test result predicts the risk of having a depressive disorder. For the YSR AFF ranges 0 to 4, 5 to 6, 7 to 9 and a score for 10 or above the following LR were found: 0.14, 0.52, 2.29 and 20.5. The nomogram in figure 6 can be used to identify the risk of a subject to suffer from a depressive episode according to ICD-10 criteria for a specific YSR AFF test result and a known pre-test risk of ICD-10 depressive episodes. For example, the prevalence of ICD-10 depressive episodes for a specific population may be p = 5% and a subject out of this population may have a YSR AFF score above 10. Thus, the risk of having a depressive episode is about 50% for this subject (as indicated by the dotted line in figure 3). Alternatively, the post-test risks can be determined by changing the probabilities in odds ratios and using the following formula: post-test odds = pretest odds × likelihood ratio.

5.5 Discussion

This study attempted to replicate and expand previous findings dealing with the diagnostic accuracy of the YSR for assessing depressive episodes in referred adolescents. Strong indication for diagnostic validity of the recently developed DSM AFF scale was found. In comparison to the CES-D, the DSM AFF scale shows quite similar results in terms of accuracy and the internal consistency of both scales showed is satisfactory.

The AUC of .907 for DSM AFF scale indicates an excellent accordance of scale and clinical disorder. AUCs as a measure of excellence for predicting diagnosis should be interpreted as follows: poor (50-.70); moderate to fair (.70-.80); good (.80-.90), and excellent (.90-1.00) (Ferdinand, 2008).

For the DSM AFF scale Ferdinand (2008) found similar results (AUC of .91) in a comparable clinical sample. However, Ferdinand's diagnoses were based on structured interviews and included adolescent information only. In the present study, diagnoses of depressive episodes

were based on clinical assessments including parent and teacher information and the DSM AFF scale was highly accurate in predicting diagnoses.

Given the high correlation between the DSM AFF and the CES-D of r = .80, similar results for the diagnostic accuracy of depression for these scales were expected. Compared to the DSM AFF scale, the CES-D performed slightly better and this difference reached statistical significance. The CES-D has been found to serve as an accurate self - rating instrument for adolescent depression in previous studies with referred (Roberts et al., 1991) and non-referred adolescents (Yang et al., 2004). The present results support these findings using the German version of the CES-D.

No multidimensional model of YSR scales was found to be superior to the DSM AFF scale. Nevertheless, the tested model resulting from multivariate logistic regression analyses in the prediction sample showed comparable predictive power to the DSM AFF scale. These results are remarkable when considering that the delinquent behavior scale was included in the model which does not relate to diagnostic criteria for ICD-10 depressive episodes. These findings can not be interpreted by comorbidity with ODD or CD because the frequency of these disorders was low in the present sample. Thus, one may assume that depressive episodes in adolescence can be accompanied by rule-braking and delinquent behavior but that these symptoms may not have exceeded subthreshold levels. Previous research has addressed this issue and has concluded that adolescent depression may overlap with other disorders (Alpert et al., 1999) and quite frequently is associated with aggressive behavior and conduct problems (Harrington, 2001a; Harrington et al., 1991; Harrington, 2001b) Shared risk factors for both conduct problems and depression such as family dysfunction could explain the overlap (Fergusson et al., 1996). However, the application of a multidimensional model from logistic regression analyses is circumstantial and not practicable for screening purposes in clinical settings. Nevertheless, the present results underline the relevance of using a multidimensional rating scale for the detection of comorbid delinquent behavior.

According to the quality efficiency indicator (dQ), raw scores of 8 and 9 were identified as the optimal cut-off scores for the YSR DSM AFF scale in the sub-samples. However, even adjacent scores were leading to reasonable results in terms of the correspondent quality efficiency. Symptoms of depressive disorders are common in the general populations of adolescents. Epidemiological studies suggest that juvenile depression is rather a continuum that is associated with problems at most levels of severity than a distinct category (Pickles et al., 2001). Thus, it is not surprising that cut-off-point analyses by quality efficiency statistics often do not lead to clear results. Furthermore, quality statistics are correcting for the rate of a positive test result and for the base rate of the disorder. However, often different costs and benefits of identifying depression including the consequences of missing a true case or erroneously identifying a subject as being affected by depression have to be taken into account. Thus, in the present study a range with acceptable cut-off scores (dQ > .60) has been defined for clinical purposes. For the DSM AFF the range of scores was from 5 to 9. Thus, a raw score of 5 shows a high sensitivity and an acceptable specificity whereas a raw score of 9 maximizes specificity and reduces sensitivity to a moderate range.

Raw scores of 8 and 11 for boys and raw scores of 10 and 14 for girls were recommended by Achenbach for borderline (T = 65) and clinical range (T = 69), respectively. Findings of the present study identified even lower values (raw score of 5 or more) as meaningful indicators of depression. The recommended borderline T-scores by Achenbach represent 7% of the most affected adolescents in the sample. However, a high number of symptoms are not necessarily leading to the diagnosis of depression. Further relevant diagnostic information about onset, duration and impairment of these symptoms is not included in the YSR DSM AFF scale. ICD-10 criteria of a mild depressive episode require only four different symptoms. Thus, in clinical settings a cut-off score of 5 is meaningful in order to reduce the number of missing subjects with depressive episodes. Further acceptable cut-off scores weighting specificity higher or equal to sensitivity measures have been provided (see Table 4). These cut-off scores are better suited for research purposes than for clinical assessments.

Some limitations of the present results should be mentioned. The samples included referred subjects with depressive episodes according to ICD-10 criteria and controls from a community sample who were comparable in terms of age and sex. There was no information on psychiatric disorders in the controls. If there had been depressive disorders in the community controls, accuracy of the scales might have been reduced. However, the main trend of the results would not have been affected by this fact.

Furthermore, because our estimates of specificity are based on a community sample, they may not generalize to mental health clinic populations where specificity is likely to be lower. Nevertheless, we recommend to consider the present cut-off scores in clinical assessments. Depressive symptoms in adolescents are common also in various psychiatric disorders other than depression like anxiety, eating disorders and attention-deficit-hyperactivity disorders (LeBlanc and Morin, 2004; O'Brien and Vincent, 2003; van Lang et al., 2006). Due to their high base rate, depressive disorders can easily be missed when testing scales and determining cut-off-scores in psychiatric samples. Therefore a lower specificity has to be taken into account and additional assessment of other psychiatric disorders has to be included in order to come to the final diagnoses.

Finally, there was no formal reliability testing of clinical ICD-10 diagnoses in the present study. In clinical settings, interviewers typically follow their initial diagnostic hypothesis by asking increasingly more specific questions in order to rule in or to rule out certain diagnoses (Doss, 2005). In the present study, diagnoses by postgraduate clinicians were based on clinical interviews with parents and children and included also teacher information. In each case, these diagnoses were confirmed by senior clinical experts so that the best estimate procedure was performed. In addition, clinicians were blind to the scores of the YSR DSM oriented scales but not blind to other YSR findings and the CES-D findings. Thus, the results based on the CES-D have to be regarded with caution because of possible criterion contamination, whereas the findings for the DSM AFF were not substantially affected.

Despite these limitations, the present findings suggest that similar to the CES-D the DSM-AFF scale of the YSR is a highly accurate screening instrument for depressive episodes according to ICD-10. Thus, the YSR as a multivariate self - rating instrument can be recommended for the diagnostic assessment of depression. For the identification of post-test risks of depressive episodes, the nomogram in figure 6 can be used. In clinical settings, a cut-off score of 5 should be considered as a starting point of the clinical assessment of depression. Additionally, further information on age of onset, continuity, impairment, specificity of symptoms and information about other psychiatric disorders should be included in order to arrive at the final diagnosis.

5.6 References

Achenbach, T.M., 1991a. Manual for the Child Behavior Check List/4-18 and 1991 Profile. Department of Psychiatry, University of Vermont, Burlington, VT.

Achenbach, T.M., 1991b. Manual for the Youth Self Report and 1991 Profile. Department of Psychiatry, University of Vermont, Burlington, VT.

Achenbach, T.M., Dumenci, L., Rescorla, L.A., 2003. DSM-oriented and empirically based approaches to constructing scales from the same item pools. J Clin Child Adolesc Psychol 32, 328-340.

Achenbach, T.M., Rescorla, L.A., 2001. Manual for the School-Age Forms and Profiles. Child Behavior Checklist. Teacher's Report Form. Youth Self-Report. An Integrated System of Multi-informant Assessment. Library of Congress, Burlington.

Alpert, J.E., Fava, M., Uebelacker, L.A., Nierenbarg, A.A., J.A., P., Worthington III, J.J., Rosenbaum, J.F., 1999. Patterns of axis I comorbidity in early-onset versus late-onset major depressive disorder. Biol Psychiatry 46, 202-211.

Angold, A., Costello, E.J., 1993. Depressive comorbidity in children and adolescents: empirical, theoretical, and methodological issues. Am J Psychiatry 150, 1779-1791.

Chorpita, B.F., Yim, L., Moffitt, C., Umemoto, L.A., Francis, S.E., 2000. Assessment of symptoms of DSM-IV anxiety and depression in children: a revised child anxiety and depression scale. Behav Res Ther 38, 835-855.

Costello, E.J., Erkanli, A., Angold, A., 2006. Is there an epidemic of child or adolescent depression? J Child Psychol Psychiatry 47, 1263-1271.

Doss, A.J., 2005. Evidence-based diagnosis: incorporating diagnostic instruments into clinical practice. J Am Acad Child Adolesc Psychiatry 44, 947-952.

Ferdinand, R.F., 2008. Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. J Anxiety Disord 22, 126-134.

Fergusson, D.M., Lynskey, M.T., Horwood, L.J., 1996. Origins of comorbidity between conduct and affective disorders. J Am Acad Child Adolesc Psychiatry 35, 451-460.

Garrison, C.Z., Addy, C.L., Jackson, K.L., McKeown, R.E., Waller, J.L., 1991. The CES-D as a screen for depression and other psychiatric disorders in adolescents. J Am Acad Child Adolesc Psychiatry 30, 636-641.

Gould, M.S., Bird, H., Jaramillo, B.S., 1993. Correspondence between statistically derived behavior problem syndromes and child psychiatric diagnoses in a community sample. J Abnorm Child Psychol 21, 287-313.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839-843.

Harrington, R., 2001a. Adolescent depression: same or different? Arch Gen Psychiatry 58, 21-22.

Harrington, R., Fudge, H., Rutter, M., Pickles, A., Hill, J., 1991. Adult outcomes of childhood and adolescent depression: II. Links with antisocial disorders. J Am Acad Child Adolesc Psychiatry 30, 434-439.

Harrington, R.C., 2001b. Childhood depression and conduct disorder: different routes to the same outcome? Arch Gen Psychiatry 58, 237-238.

Hautzinger, M., Bailer, M., 1993. Allgemeine Depressions-Skala (ADS). Deutsche Form der " Center of Epidmiologic Studies Depression Scale" (CES-D). Beltz, Weinheim.

Kovacs, M., Gatsonis, C., Paulauskas, S.L., Richards, C., 1989. Depressive disorders in childhood. IV. A longitudinal study of comorbidity with and risk for anxiety disorders. Arch Gen Psychiatry 46, 776-782.

Kovacs, M., Paulauskas, S., Gatsonis, C., Richards, C., 1988. Depressive disorders in childhood. III. A longitudinal study of comorbidity with and risk for conduct disorders. J Affect Disord 15, 205-217.

Kraemer, H.C., 1992. Evaluating medical tests. Objective and quantitative guidelines. Sage Publications, Inc., Newbury Park.

Krol, N.P., De Bruyn, E.E., van Aarle, E.J., Van den Bercken, J., 2001. Computerized screening for DSM classificiations using CBCL/YSR extended checklist: A clinical tryout. Comput Hum Behav 17, 315-337.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

LeBlanc, N., Morin, D., 2004. Depressive symptoms and associated factors in children with attention deficit hyperactivity disorder. J Child Adolesc Psychiatr Nurs 17, 49-55.

Lengua, L.J., Sadowski, C.A., Friedrich, W.N., Fisher, J., 2001. Rationally and empirically derived dimensions of children's symptomatology: expert ratings and confirmatory factor analyses of the CBCL. J Consult Clin Psychol 69, 683-698.

Lewinsohn, P.M., Hops, H., Roberts, R.E., Seeley, J.R., Andrews, J.A., 1993. Adolescent psychopathology: I. Prevalence and incidence of depression and other DSM-III-R disorders in high school students. J Abnorm Psychol 102, 133-144.

Morgan, C.J., Cauce, A.M., 1999. Predicting DSM-III-R disorders from the Youth Self-Report: analysis of data from a field study. J Am Acad Child Adolesc Psychiatry 38, 1237-1245.

O'Brien, K.M., Vincent, N.K., 2003. Psychiatric comorbidity in anorexia and bulimia nervosa: nature, prevalence, and causal relationships. Clin Psychol Rev 23, 57-74.

Pickles, A., Rowe, R., Simonoff, E., Foley, D., Rutter, M., Silberg, J., 2001. Child psychiatric symptoms and psychosocial impairment: relationship and prognostic significance. Br J Psychiatry 179, 230-235.

Radloff, L.S., 1977. The CES-D scale: a self report depression scale for research in general populations. Appl. Psychol. Meas. 1, 385-401.

Roberts, R.E., Lewinsohn, P.M., Seeley, J.R., 1991. Screening for adolescent depression: a comparison of depression scales. J Am Acad Child Adolesc Psychiatry 30, 58-66.

Roberts, R.E., Roberts, C.R., Chen, I.G., 2000. Fatalism and risk of adolescent depression. Psychiatry 63, 239-252.

Roberts, R.E., Vernon, S.W., 1983. The Center for Epidemiologic Studies Depression Scale: its use in a community sample. Am J Psychiatry 140, 41-46.

Rohde, P., Jorgensen, J.S., Seeley, J.R., Mace, D.E., 2004. Pilot evaluation of the Coping Course: a cognitive-behavioral intervention to enhance coping skills in incarcerated youth. J Am Acad Child Adolesc Psychiatry 43, 669-676.

Saxena, K., Silverman, M.A., Chang, K., Khanzode, L., Steiner, H., 2005. Baseline predictors of response to divalproex in conduct disorder. J Clin Psychiatry 66, 1541-1548.

Shaffer, D., Fisher, P., Dulcan, M.K., Davies, M., Piacentini, J., Schwab-Stone, M.E., Lahey, B.B., Bourdon, K., Jensen, P.S., Bird, H.R., Canino, G., Regier, D.A., 1996. The NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA Study. Methods for the Epidemiology of Child and Adolescent Mental Disorders Study. J Am Acad Child Adolesc Psychiatry 35, 865-877.

Silverman, W.K., Saavedra, L.M., Pina, A.A., 2001. Test-retest reliability of anxiety symptoms and diagnoses with the Anxiety Disorders Interview Schedule for DSM-IV: child and parent versions. J Am Acad Child Adolesc Psychiatry 40, 937-944.

Steinhausen, H.C., Haslimeier, C., Winkler Metzke, C., 2006. The outcome of episodic versus persistent adolescent depression in young adulthood. J Affect Disord 96, 49-57.

Steinhausen, H.C., Metzke, C.W., Meier, M., Kannenberg, R., 1998. Prevalence of child and adolescent psychiatric disorders: the Zurich Epidemiological Study. Acta Psychiatrica Scandinavica 98, 262-271.

Steinhausen, H.C., Winkler Metzke, C., 2000. [The General Depression Scale in diagnosis of adolescents]. Prax Kinderpsychol Kinderpsychiatr 49, 419-434.

Steinhausen, H.C., Winkler Metzke, C., Kannenberg, R., 1996. Handbuch: Elternfragebogen über das Verhalten von Kindern und Jugendlichen. Die Zürcher Ergebnisse zur deutschen Fassung der Child Behavior Checklist (CBCL). Zentrum für Kinder- und Jugendpsychiatrie der Universität Zürich, Zürich.

Steinhausen, H.C., Winkler Metzke, C., Meier, M., Kannenberg, R., 1997. Behavioral and emotional problems reported by parents for ages 6 to 17 in a Swiss epidemiological study. Eur Child Adolesc Psychiatry 6, 136-141.

van Lang, N.D., Ferdinand, R.F., Oldehinkel, A.J., Ormel, J., Verhulst, F.C., 2005. Concurrent validity of the DSM-IV scales Affective Problems and Anxiety Problems of the Youth Self-Report. Behav Res Ther 43, 1485-1494.

van Lang, N.D., Ferdinand, R.F., Ormel, J., Verhulst, F.C., 2006. Latent class analysis of anxiety and depressive symptoms of the Youth Self-Report in a general population sample of young adolescents. Behav Res Ther 44, 849-860.

Vreugdenhil, C., van den Brink, W., Ferdinand, R., Wouters, L., Doreleijers, T., 2006. The ability of YSR scales to predict DSM/DISC-C psychiatric disorders among incarcerated male adolescents. Eur Child Adolesc Psychiatry 15, 88-96.

Weinstein, S.R., Noam, G.G., Grimes, K., Stone, K., Schwab-Stone, M., 1990. Convergence of DSM-III diagnoses and self-reported symptoms in child and adolescent inpatients. J Am Acad Child Adolesc Psychiatry 29, 627-634.

Weissman, M.M., Sholomskas, D., Pottenger, M., Prusoff, B.A., Locke, B.Z., 1977. Assessing depressive symptoms in five psychiatric populations: a validation study. Am J Epidemiol 106, 203-214.

Yang, H.J., Soong, W.T., Kuo, P.H., Chang, H.L., Chen, W.J., 2004. Using the CES-D in a two-phase survey for depressive disorders among nonreferred adolescents in Taipei: a stratum-specific likelihood ratio analysis. J Affect Disord 82, 419-430.

6 Study 3: Predictability and construct validity of oppositional defiant disorder in children and adolescents with ADHD combined type[3]

6.1 Abstract

Three recently identified dimensions of Oppositional Defiant Disorders (ODD) might indicate different pathways of future emotional and behavioral problems. The present study aimed at testing the construct validity of these three dimensions of ODD and the diagnostic accuracy of two common parent rating scales in predicting ODD and the dimensions of ODD in a large referred sample of children and adolescents with Attention-Deficit-Hyperactivity Disorder (ADHD) combined type. Subjects came from the International Multicentre ADHD Genetics (IMAGE) Study. Receiver operating characteristic (ROC) analyses showed adequate diagnostic accuracy of the Conners' parent rating scale revised (CPRS-R) and the parent version of the strength and difficulties questionnaire (PSDQ) in predicting ODD in this ADHD sample. The three factor structure of ODD was confirmed by confirmatory factor analyses. The CPRS-R emotional lability scale significantly predicted the ODD irritable dimension, that is associated with severe mood dysregulation in ADHD referred youth.

Keywords: Oppositional-Defiant Disorder; Attention-Deficit-Hyperactivity Disorder; Conners' Parent Rating Scale Revised; Strength and Difficulties Questionnaire; Irritability; Emotional lability.

---

[3] Aebi, M., Müller, U. C., Asherson, P., Banaschewski, T., Buitelaar, J., Ebstein, R., Eisenberg, J., Gill, M., Manor, I., Miranda, A., Oades, R. D., Roeyers, H., Rothenberger, A., Sergeant, J., Sonuga-Barke, E., Thompson, M., Taylor, E., Faraone, S. V., and Steinhausen, H.-C. (submitted) Predictability and construct validity of oppositional defiant disorder in children and adolescents with ADHD combined type

6.2 Introduction

High rates of co-morbid oppositional defiant disorder (ODD) and conduct disorder (CD) have been found in subjects with ADHD (e.g. Angold, Costello, & Erkanli, 1999) and milder forms of conduct problems like ODD are strongly related to ADHD symptoms (Christiansen et al., 2008). Recent findings support the idea that the development of later conduct disorders in subjects with ADHD is mediated by co-morbid ODD (Biederman, Petty et al., 2008; Burke, Loeber, Lahey, & Rathouz, 2005; Lahey, Loeber, Burke, Rathouz, & McBurnett, 2002; van Lier, van der Ende, Koot, & Verhulst, 2007). Furthermore, ODD seems to be a pivotal disorder for the development of conduct, affective and anxiety disorders in youth (Burke et al., 2005; Nock, Kazdin, Hiripi, & Kessler, 2007).

In mental health clinics, the diagnosis of ADHD and ODD in children and adolescents largely rests on detailed interviews with their parents and caretakers. In addition, parent and teacher rating scales like the Conners' Parent (CPRS; Conners, Sitarenios, Parker, & Epstein, 1998a) and Teacher Rating Scale (CTRS; Conners, Sitarenios, Parker, & Epstein, 1998b) or the Strength and Difficulties Questionnaire (SDQ; Goodman, 1997, 2001) contribute considerable information to the assessment process. Besides the narrowband syndrome scale of attention problems and hyperactivity, these instruments also include specific scales to screen for ODD (Conners, 1997; Goodman, 2001; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000).

The Conners' Parent Rating Scale (CPRS) and related versions of the CPRS have been used in previous studies as screening instruments for various mental disorders and as outcome parameters in treatment studies dealing with externalizing behavior problems including ADHD (for an overview see Gianarris, Golden, & Greene, 2001). Although the CPRS-R has been widely used in clinical and research settings, some quite fundamental criticisms have been raised which primarily deal with the suitability of subscales measuring

problems other than ADHD and, particularly, oppositional problems (Collett, Ohan, & Myers, 2003).

In comparison to the CPRS-R, the SDQ is of more recent origin and is a shorter instrument for screening the most important mental disorders in childhood and adolescence. The SDQ addresses 5 narrowband syndromes: emotional symptoms, conduct problems, hyperactivity, peer problems and pro-social behavior. A computer algorithm has been developed for the prediction of oppositional-conduct, hyperactive-inattention, anxious-depressed or any psychiatric disorder. The predictions from the computer algorithm of the multi-informant SDQ has been found to correlate with clinical diagnoses of CD/ODD in referred subjects from Europe, Bangladesh and Australia (Goodman, Renfrew, & Mullick, 2000; Mathai, Anderson, & Bourne, 2004). High sensitivity in the detection of clinical CD/ODD has been established (86-93%) whereas specificity was only modest indicating that the SDQ was over-including subjects in these samples. On the other hand, in a community sample, a smaller number of subjects (68.2%) with internet-interview based diagnosis of CD/ODD (DAWBA; Goodman, Ford, Richards, Gatward, & Meltzer, 2000) were rated as having a probable diagnosis of CD/ODD based on the SDQ (Goodman, Ford, Simmons et al., 2000). Due to the high rate of false positives, the SDQ seems to be more suitable for screening rather than for establishing diagnoses in community samples.

Only until recently, evidence has been missing that in contrast to other rating scales (e.g. the Child Behavior Checklist (CBCL); Biederman, Ball, Monuteaux, Kaiser, & Faraone, 2008; Eiraldi, Power, Karustis, & Goldstein, 2000) both the parent SDQ (PSDQ) and the CPRS-R also predict ODD in ADHD subjects. Furthermore, the CPRS-R oppositional scale (CPRS-R OPP) has never been specifically tested as regards its predictive validity for ODD. A recent study based on the IMAGE sample has analyzed these scales in the identification of conduct problems (Christiansen et al., 2008). This study found that the CPRS-R OPP and the PSDQ conduct problem scales (PSDQ CP) yielded the best discrimination of pure ADHD, ODD and CD. Additional ROC analyses confirmed adequate diagnostic accuracy in the prediction of

CD and found a cut-off-score above the 85th percentile as best discriminator for both scales. However, the prediction of ODD as a separate disorder apart from CD has not yet been analyzed in this study.

Therefore, as the first step for the present study we aimed to assess the predictive validity of the CPRS-R and the PSDQ in the prediction of ODD taking previous findings into account that confirmed ODD as a discrete psychiatric disorder regarding impairment and co-morbidity (Burke et al., 2005; Greene et al., 2002). Furthermore, cut-off analyses will be performed by quality efficiency statistics and the results of the PSDQ will be compared to the results of the proposed computer algorithm of the SDQ.

Different dimensions of ODD may by important regarding course and co-morbidity. The development of later emotional disorders may be predicted by the affective features in ODD symptoms reflecting negative and temperamental qualities (e.g. 'often angry and resentful' 'temper tantrums') (Burke et al., 2005). Recently, Stringaris and Goodman (in press) defined three a priori dimensions of oppositionality which were labeled ODD-irritable, ODD-headstrong and ODD-hurtful based on the DSM-IV criteria for ODD. The authors found different associations with other disorders in a large community sample of youth aged 5 to 16 years using parent and teacher information from a structured internet based diagnostic interview (Development and Well-Being Assessment; DAWBA) (Goodman, Ford, Richards et al., 2000). The ODD-irritable dimension was related to emotional disorders, whereas the ODD-headstrong dimension was related to ADHD and all three dimensions were related to conduct disorder. The authors concluded that these three dimensions may be important predictors of the aetiology, prognosis and treatment of ODD.

Therefore, the second aim of the present study was to test the construct validity of these three dimensions in a sample including ODD subjects. In contrast to Stringaris and Goodman (in press), the item "often deliberately annoys people" was assigned to the ODD-hurtful dimension because in a previous study this item was most strongly correlated with spiteful

behavior (Speltz, McClellan, DeKlyen, & Jones, 1999). In a final step, the accuracy of the CPRS-R and the PSDQ in addressing these separate dimensions was tested in subjects with and without ODD.

6.3 Methods

6.3.1 Participants

The IMAGE study comprises 3229 offspring from 1187 fathers and 1341 mothers. Probands participating in the present study were European Caucasians aged 5-17 years that had been recruited in 12 child and adolescent psychiatry clinics representing eight countries: Belgium, Germany, Switzerland, Holland, Ireland, Israel, Spain and United Kingdom. Entry criteria for probands were a clinical diagnosis of ADHD based on DSM-IV criteria and access to one or both biological parents and one or more full siblings for DNA collection and clinical assessment. Exclusion criteria applying to both probands and siblings included autism, epilepsy, IQ < 70, brain disorders and any genetic or medical disorder associated with externalizing behaviors that might mimic ADHD.

The original sample of 1401 probands has been restricted to 1225 subjects with ADHD combined type. Furthermore 91 (7%) were excluded due to missing information on DSM-IV ODD criteria and another 31 (3%) subjects due to more than 10% missing items in the CPRS-R or the PSDQ. Thus, the final sample consisted of 1093 probands with a mean age of 10.8 years (SD 2.8 years). 956 subjects were male (87.5%) and 726 (66.4%) subjects from the present sample fulfilled DSM-IV criteria of ODD based on the PACS-interview (see below).

6.3.2 Measures

Diagnoses were based on a standardized, semi-structured interview with the parents (Parental Account of Childhood Symptoms, [PACS]; W. Chen & Taylor, 2006; Taylor,

Schachar, Thorley, & Wieselberg, 1986) that includes four sections: hyperactivity (attention span, fidgetiness and restlessness), defiance (e.g., tantrums, disobedience and destructiveness), emotionality (e.g., misery, worries, fears) and comorbid disorders (autistic spectrum, attachment, mania, substance-abuse, psychotic symptoms, obsessive-compulsive symptoms, and other specific developmental and neurological conditions). The diagnoses of ADHD, ODD and CD were based on an algorithm which is appropriate for symptom count, age, time interval and impairment according to DSM-IV criteria. All subjects from the present sample were referred for ADHD combined type.

The long form of the revised Conners' Parent Rating Scale (CPRS-R: L) consisting of 80 items was used in the present study. The CPRS-R is a reliable, accurate, and relatively brief measure of parental perceptions of children's disruptive behavior. Adequate psychometric properties have been confirmed (Conners, 1997; Conners et al., 1998a) The seven syndrome scales (Cognitive Problems, Oppositional, Hyperactivity-Impulsivity, Anxious-Shy, Perfectionism, Social Problems and Psychosomatics), the ADHD index and the two sub-scales of the Conners'-Global Index (CGI; restless-impulsive, emotional lability) were included in the present study.

The SDQ is a brief behavioral screening questionnaire for 4 to 16 year olds. There are versions for adolescents (starting from 11 years onwards), parents and teachers. The SDQ consists of five syndrome scales (emotional symptoms, conduct problems, hyperactivity, peer problems and pro-social behavior) and can be obtained free via the internet (http://www.sdqinfo.com). Adequate psychometric properties of the scales have been documented (Goodman, 1997, 2001).

6.3.3 Analytic procedure

To study the diagnostic accuracy in the prediction of ODD, ROC analyses were performed separately for each CPRS-R syndrome scale including the two CGI subscales and the ADHD

index scale. Furthermore, the PSDQ scales were included in the ROC analyses. The pro-social behavior scale was excluded because it does not address problem behavior. To compare different scales within the same sample, a critical z-ratio was calculated using a formula correcting for the non-independence of the scales (Hanley & McNeil, 1983). Finally, the optimal cut-off-score for the best scales was established: Efficiency (EFF) was calculated by the sum of true positives (TP) and true negatives (TN). In order to correct EFF for independence of the base rate (P) in the sample and to take into account the rate of a positive test result (Q), a quality index of efficiency was calculated using the following formula: $dQ = [EFF – PQ - (1 - P)(1 - Q)]/[1 - PQ- (1 - P)(1 - Q)]$ (Kraemer, 1992). In addition, the proposed computer algorithm for the identification of possible and probable CD/ODD cases was compared to the results based on the cut-off-score analyses.

Construct validity of the three ODD dimensions was analyzed by use of confirmatory factor analysis including all symptoms accounting for ODD in the PACS. Each symptom was rated as present or absent according to the corresponding PACS algorithm. Maximum likelihood statistics using the AMOS 16 software were used to assess three different recommended goodness of fit indicators (GFI) (Hair, Black, Babin, Anderson, & Tatham, 2006), i.e., the root mean square residual (RMR) as indicator of the unexplained co-variances of the model, the root mean square error of approximation (RMSEA) which includes a parsimony correction, and the comparative fit index (CFI) for evaluating the hypothesized model compared to a null model. Acceptance of any model was based on the following cut-offs: $RMR < 0.05$, $RMSEA < 0.08$ and $CFI > 0.95$ (Hu & Bentler, 1999; Marsh, Kit-Tai, & Zhonglin, 2004).

In a further step, the prediction of the three factor structure of ODD by the CPRS-R and the PSDQ was analyzed. Backward linear regression analyses were performed including all syndrome and index scales of the CPRS-R and the PSDQ of subjects both with and without ODD. For these analyses, the total sample was split into two subgroups each, namely, prediction and cross-validation sub-samples. Group assignment was done by random sampling controlling for ODD, age and sex. When using exploratory analyses like backward

linear regression, cross-validation can be a helpful technique in order not to over-interpret results in terms of generalizability (Leon, Olfson, Weissman, Portera, & Sheehan, 1996).

6.4 Results

Means and standard deviations of the CPRS-R scores and the PSDQ scores are shown in Table 11. Internal consistency as measured by Cronbach's alpha was .88 for the CPRS-R oppositional scale and .66 for the PSDQ CP. The scores of the two scales were strongly correlated ($r = .67$, $p < 0.001$).

Table 11. *Means and standard deviations (raw scores) of CPRS-R and the PSDQ separate for subjects with and without co-morbid ODD in the entire sample, in the prediction sample and the cross-validation sub-sample*

Sample	Entire sample (N = 1093)				Prediction sub-sample (N = 546)				Cross-validation sub-sample (N =547)			
	ODD (N = 726)		no ODD (N = 367)		ODD (N = 363)		no ODD (N = 183)		ODD (N = 363)		no ODD (N = 184)	
	means	SD	means	SD	means	SD	means	SD	means	SD	means	SD
Age	10.83	2.71	10.65	2.83	10.86	2.61	10.61	2.96	10.80	2.81	10.68	2.71
CPRS-R Syndrome Scales												
Oppositional	19.41	5.89	13.06	6.49	19.63	5.70	13.05	6.35	19.20	6.07	13.07	6.65
Cognitive Problems / Inattention	24.70	6.51	23.38	6.73	25.05	6.27	23.13	6.01	24.35	6.74	23.64	7.38
Hyperactivity	17.76	5.12	16.10	5.84	17.91	4.98	16.14	5.53	17.62	5.26	16.05	6.16
Anxious-Shy	6.59	5.12	4.73	4.55	6.80	5.12	4.72	4.75	6.38	5.13	4.74	4.36
Perfectionism	6.29	4.67	5.06	4.26	6.27	4.68	4.93	4.21	6.31	4.66	5.18	4.31
Social Problems	6.10	4.02	4.37	3.54	6.09	4.03	4.36	3.42	6.10	4.01	4.39	3.66
Psychosomatic	4.43	3.98	3.22	3.43	4.63	3.95	3.04	3.20	4.23	4.02	3.39	3.64
ADHD Index	27.72	5.67	25.83	6.35	27.91	5.64	25.74	5.75	27.53	5.70	25.91	6.91
CGI: Restless-Impulsive	16.21	3.41	14.25	4.03	16.21	3.43	14.44	3.73	16.21	3.40	14.06	4.30
CGI: Emotional Lability	5.28	2.16	3.55	2.36	5.45	2.17	3.55	2.30	5.11	2.14	3.55	2.42
PSDQ Scales												
Emotional Symptoms	4.16	2.51	3.25	2.43	4.22	2.46	3.25	2.46	4.08	2.57	3.22	2.40
Conduct Problems	5.34	2.18	3.43	2.17	5.47	2.15	3.38	2.04	5.21	2.21	3.47	2.29
Hyperactivity	8.58	1.56	8.31	1.82	8.69	1.53	8.43	1.76	8.47	1.58	8.18	1.88
Peer Problems	4.32	2.60	3.37	2.57	4.23	2.57	3.42	2.51	4.42	2.64	3.32	2.65

Table 12 shows the results of the ROC - analyses for all CPRS-R syndrome scales and the PSDQ scales for predicting ODD. The CPRS-R oppositional scale showed the best prediction (AUC = .77) in contrast to all remaining CPRS-R scales. The PSDQ CP showed the best prediction (AUC = .73) in contrast to the remaining SDQ problem scales. The CPRS-R oppositional scale was superior when compared to the SDQ CP scale (z = 2.248, p = 0.014). There were no gender differences in the prediction of ODD by the CPRS-R OPP (boys AUC = .76; girls AUC = .79; z = -.63, p = 0.263) and for the PSDQ CP (boys AUC = .73; girls AUC = .75; z = -.34, p = 0.367).

Table 12. *ROC analysis findings with area under the curve (AUC) of the CPRS-R and the PSDQ problem syndrome scales*

Sample (N = 1093)	**AUC**	**SE**	**p**
CPRS-R problem syndrome scales			**Deviation from CPRS-R Oppositional**
Oppositional	.77	.015	--
Cognitive Problems / Inattention	.56	.018	< 0.001
Hyperactivity	.58	.018	< 0.001
Anxious-Shy	.61	.018	< 0.001
Perfectionism	.58	.018	< 0.001
Social Problems	.63	.018	< 0.001
Psychosomatic	.59	.018	< 0.001
ADHD Index	.59	.018	< 0.001
CGI: Restless-Impulsive	.64	.018	< 0.001
CGI: Emotional Lability	.71	.017	< 0.001
PSDQ problem syndrome scales			**Deviation from PSDQ Conduct Problems**
Emotional Symptoms	.61	.018	< 0.001
Conduct Problems	.73	.016	--
Hyperactivity	.53	.019	< 0.001
Peer Problems	.61	.018	< 0.001

Note. All scales showed significant deviance of AUC from random prediction (AUC = .5) except the PSDQ hyperactivity scale (p=0.07).

The results of the cut-off analyses are shown in table 13. For the CPRS-R OPP, a cut-off-score of 15 to 16 was established based on the efficiency statistics (dQ = .40). Overall 73% of the subjects were classified correctly by this score. Sensitivity, specificity and positive and negative predictive power ranged between .58 and .80. For the PSDQ CP, the optimal cut-off-score was 5 (dQ = .34). The corresponding sensitivity and specificity scores were in a similar range between .55 and .79. In addition, the point-biserial correlation coefficients were .44 ($p < 0.001$) between ODD and CPRS-R OPP, and .38 ($p < 0.001$) between ODD and the PSDQ CP.

Table 13. *Cut-off-score analyses of the CPRS-R oppositional scale and the PSDQ CP by a quality efficiency indicator (d_Q)*

Cut-off-score/ Computer algorithm	Base rates	SE	SP	PPP	NPP	EFF	d_Q	LR +	LR-
CPRS-R oppositional scale									
15	0.66	0.79	0.61	0.80	0.60	0.73	0.40	0.10	0.39
16	0.62	0.76	0.65	0.81	0.58	0.72	0.40	0.08	0.38
Parent SDQ conduct problem scale									
4	0.68	0.79	0.55	0.78	0.57	0.71	0.34	0.11	0.35
PSDQ computer algorithm for CD/ODD									
Possible CD/ODD disorder	0.68	0.79	0.55	0.78	0.57	0.71	0.34	0.11	0.35
Probable CD/ODD disorder	0.49	0.61	0.75	0.83	0.50	0.66	0.32	0.03	0.27

Note. SP = specificity; SE = sensitivity; PPP = positive predictive power; NPP = negative predictive power; EFF = efficiency; d_Q = quality index for efficiency; LR+ = likelihood ratio of a positive test; LR- = likelihood ratio of a negative test.

As can be seen from table 13, the proposed computer algorithm for the SDQ in predicting possible CD/ODD resulted in exactly the same results as the quality efficiency approach. Finally, the corresponding computer algorithm for probable CD/ODD, which considers the social impact of the symptoms, showed quite comparable efficiency with a reduced sensitivity score when compared to the specificity score.

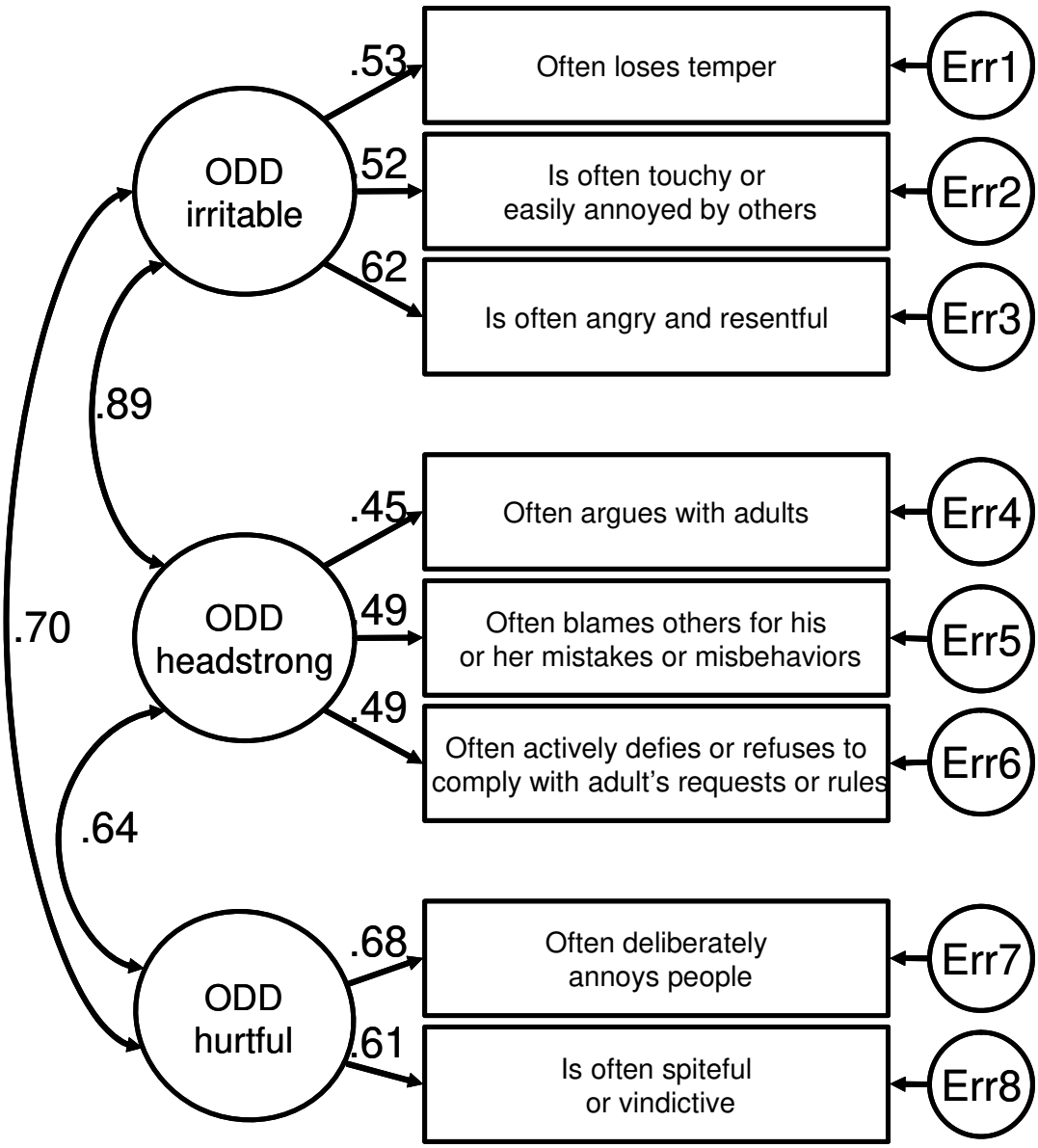

Figure 7 Confirmatory factor analysis of the 8 DSM-IV ODD criteria. Standardized regression weights and correlations between the three ODD factors ODD-Irritable, ODD-Headstrong and ODD-Hurtful.

In the second part of the analyses, the three-factor-structure of the ODD was tested by confirmatory factor analysis with maximum likelihood estimation of model parameters. The factor structure and parameter estimates are shown in figure 7. All three goodness of fit indicators suggested that the model had an excellent fit to the data (RMR =.005, RMSEA = .039 and CFI = .976). The three factors were highly correlated as shown in figure 1. However, compared to the three factor solution a single factor model of ODD showed a decreased fit to the present data (RMR =.008, RMSEA = .065 and CFI = .921).

Finally, backward linear regression analyses (probability level of F for entry = .001 and for removal = .01) separately for subjects with ODD (N = 726) and without ODD (N = 367) were performed in a prediction sub-sample (ODD: N= 363, non-ODD: N= 183) and cross-validated in a further sub-sample (ODD: N= 363, non-ODD: N= 184). The results for the prediction of ODD-irritable, ODD-headstrong and ODD-hurtful are shown in Table 14. All tested regression models were highly significant. The CPRS-R emotional lability scale (CPRS-R EL) significantly predicted ODD-irritable for subjects who did not fulfill criteria for ODD. A multivariate model including the CPRS-R EL and the CPRS-R OPP was found to significantly predict ODD-Irritable in subjects who fulfilled criteria for ODD. Both of these prediction models were confirmed in the cross-validation sub-sample as indicated by the comparable R-values ranging from .31 to .35. In the combined sample of subjects with and without ODD the correlations between CPRS-R EL and ODD-irritable amounted to r = .42 in the prediction sub-sample and r = .48 in the cross-validation sub-sample. For the ODD-headstrong dimension, no specific model resulting from backward regression analyses was confirmed in the cross-validation sample. This was true for both the ODD and the non-ODD condition. Finally, only the CPRS-R oppositional scale was found to predict the ODD-hurtful dimension in ODD (R = .27) and non ODD (R = .35) subjects. However, only in subjects with ODD (R = .31) but not without ODD (R = .10) the prediction model was confirmed in the corresponding cross-validation sample.

Table 14. *Prediction of ODD-dimensions by the PSDQ problem scales, the CPRS-R problem and index scales based on backward linear regression analyses in the prediction sample separate for subjects with and without ODD*

Prediction model	Model summary		ANOVA			Coefficients		
	R (prediction sample)	R (cross-validation sample)	Df	F	Sign.	Beta	T	Sign.
ODD diagnosis								
ODD-Irritable	.345	.326	2	24.29	.000			
CPRS-R oppositional behavior						.179	2.77	.006
CPRS-R CGI emotional lability						.201	3.11	.002
ODD-Headstrong	.261	.125	2	13.18	.000			
CPRS-R oppositional behavior						.153	2.77	.006
CPRS-R ADHD Index						.159	2.87	.004
ODD-Hurtful	.268	.314	1	28.03	.000			
CPRS-R oppositional behavior						.268	5.29	.000
No ODD diagnosis								
ODD-Irritable	.340	.311	1	23.71	.000			
CPRS-R CGI emotional lability						.340	4.87	.001
ODD-Headstrong	.377	.239	2	14.91	.000			
CPRS-R oppositional behavior						.439	5.55	.001
CPRS-R CGI restless impulsive						-2.12	-2.63	.009
ODD-Hurtful	.348	.097	1	25.01	.000			
CPRS-R oppositional behavior						.348	5.00	.000

Note. Beta = standardized regression coefficent. Prediction sample with ODD: N = 363; cross-validation sample with ODD: N = 363; prediction sample without ODD: N = 183; cross-validation sample without ODD: N = 184.

6.5 Discussion

The first part of the present study dealt with testing the diagnostic accuracy of two common parent rating scales for predicting ODD in a sample of ADHD referred youth. Construct

validity for three previous described dimensions of ODD were analyzed in the second part. Finally, the diagnostic accuracy of the CPRS-R and p–SDQ in the prediction of these three dimensions of ODD was examined.

Diagnostic accuracy was tested by ROC leading to the calculation of the AUC. This measure of excellence in the prediction of diagnoses should be interpreted as follows: poor (50-.70); moderate to fair (.70-.80); good (.80-.90), and excellent (.90-1.00). Accordingly, the AUCs for CPRS-R OPP (.77) and PSDQ CP (.73) indicate an acceptable convergence of these scales with the diagnosis of ODD. These results are quite comparable with the diagnostic accuracy of the CBCL aggressive behavior scale in a pure ADHD sample (Biederman, Ball et al., 2008) and in a mixed ADHD sample with unreferred controls (Hudziak, Copeland, Stanger, & Wadsworth, 2004).

In comparison to the present findings, higher AUCs based on parental ratings have been reported in the prediction of various psychiatric disorders other than ODD, e.g. for obsessive compulsive disorders (Hudziak et al., 2006) and for ADHD (W. J. Chen, Faraone, Biederman, & Tsuang, 1994). Furthermore, a better diagnostic accuracy has been found also in the in study by Christiansen et al. (2008) in the prediction of CD in ADHD subjects by the PSDQ CP and the CPRS-R OPP in a smaller subsample of the IMAGE study. The differences in diagnostic accuracy may be partly due to sample and rater effects. For instance, parent ratings of ODD and ADHD have been found to be biased by observer characteristics such as depressed mood and levels of stress (van der Oord, Prins, Oosterlaan, & Emmelkamp, 2006). Thus, parents under stress or with depressed mood may experience ODD symptoms as particular aversive. However, whether or not the relationship with a child showing ODD is even more aversive for parents and may even more negatively influence diagnostic accuracy of ODD than in pure ADHD still has to be shown.

In contrast to the present findings, Biederman and colleagues (2008) in their prediction of ODD in ADHD subjects by use of the CBCL found higher AUCs and efficiencies in girls than in boys. These results may be due to using standardized T-scores rather than raw scores.

In the present study, a cut-off-score of 15/16 on the CPRS-R oppositional problem scale and a cut-off-score of 4 on the PSDQ CP in the detection of ODD were found by quality efficiency statistics. For the CPRS-R, raw scores of 15/16 correspond to T-scores of 66-73 in boys and to 70-75 in girls. On the other hand a cut-off-score of T = 65 has been recommended for screening for ODD (Conners, 1997). Whereas this lower cut-off-score may be accurate in clinical settings the same score will be over- inclusive in an ADHD sample and particular for girls. However, the PSDQ computer algorithm for possible ODD/CD seems to work well in subjects with or without comorbid ADHD.

Whereas a recent study by Stringaris and Goodman (in press) focussed on the predictive validity of three theoretical established dimensions of ODD, the present study addressed the construct validity of these dimensions. By replicating the findings by Stringaris and Goodman (in press), the present study serves as a cross-validation of the three ODD dimensions labelled ODD-irritable, ODD-headstrong and ODD-hurtful. The present results indicate that these three factors are valid and meaningful dimensions also in subjects referred for ADHD with comorbid ODD. Furthermore, our results convincingly show that a three factor structure of ODD is more appropriate than a single general factor of ODD. This finding may have nosological implications for the upcoming DSM-V criteria. Currently, no definite conclusions reflecting the usefulness of these ODD-dimensions regarding aetiology, treatment and prediction of future disorders can be made. However, recent results showed different relations of these ODD dimensions to co-occurring disorders and suggest meaningful implication for clinical practice (Stringaris & Goodman, in press).

Finally, potential predictors of these three dimensions were analyzed. Whereas the prediction of ODD-headstrong and ODD-hurtful by the CPRS-R and the PSDQ led only to ambiguous

results, except for the CPRS-R OPP scale, the CPRS-R EL is a meaningful predictor of the irritable dimension of ODD. Furthermore, the CPRS-R EL predicted ODD irritability also in subjects with no ODD indicating that this dimension is also important in pure ADHD subjects. Despite the fact that the CPRS-R EL scale consists only of three items (i.e. temper outbursts, crying, mood changes), this scale is rather sensitive in predicting ODD-irritable as indicated by correlations ranging between r = .421 and r = .479. Thus, the predictive validity of the CPRS-R EL originally found in both exploratory and confirmatory factor analyses (Parker, Sitarenios, & Conners, 1996) was confirmed by the present results.

Recently, the role of irritability in ADHD with comorbid ODD has been addressed in the context of severe mood dysregulation (SMD; Carlson, 2007). Next to abnormal mood, the diagnostic criteria of SMD include symptoms which are similar to ADHD (e.g. distractibility, pressured speech) and a markedly increased reactivity to negative emotional stimuli (similar to ODD-irritable). Furthermore, Waschbusch et al. (2002) found increased anger expression and increased heart rate after mild provocation in a sample that was comorbid for ADHD/ODD but not in ADHD or ODD only subjects. Thus, the present results indicate that the construct of SMD is related to the ODD-Irritable dimension in ADHD subjects and the CPRS-R EL may be also used as an initial screening instrument in the prediction of irritability.

Some limitations of the present findings have to be mentioned. First, the present results were based on a referred ADHD sample and may not generalize to other community and clinical samples with different base rates and characteristics of ODD. Secondly, the present findings are based on parental ratings of ODD. Multi-informant diagnostic criteria might shed further light on the prediction of these ODD dimensions.

6.6 References

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. Journal of Child Psychology and Psychiatry, 40(1), 57-87.

Biederman, J., Ball, S. W., Monuteaux, M. C., Kaiser, R., & Faraone, S. V. (2008). CBCL clinical scales discriminate ADHD youth with structured-interview derived diagnosis of oppositional defiant disorder (ODD). Journal of Attention Disorders, 12(1), 76-82.

Biederman, J., Petty, C. R., Dolan, C., Hughes, S., Mick, E., Monuteaux, M. C., et al. (2008). The long-term longitudinal course of oppositional defiant disorder and conduct disorder in ADHD boys: findings from a controlled 10-year prospective longitudinal follow-up study. Psychological Medicine, 38(7), 1027-1036.

Burke, J. D., Loeber, R., Lahey, B. B., & Rathouz, P. J. (2005). Developmental transitions among affective and behavioral disorders in adolescent boys. Journal of Child Psychology and Psychiatry, 46(11), 1200-1210.

Carlson, G. A. (2007). Who are the children with severe mood dysregulation, a.k.a. "rages"? American Journal of Psychiatry, 164(8), 1140-1142.

Chen, W., & Taylor, E. (2006). Parental account of children's symptoms (PACS), ADHD phenotypes and it's application to melecular genetic studies. In R. D. Oades (Ed.), Attention-deficit/hyperactivity disorder and the hyperkinetic syndrome: current ideas and ways forward (pp. 3-20). Hauppauge NY: Nova Science Publishing Inc.

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist scales for attention-deficit hyperactivity disorder: a receiver-operating characteristic analysis. Journal of Consulting and Clinical Psychology, 62, 1017-1025.

Christiansen, H., Chen, W., Oades, R. D., Asherson, P., Taylor, E. A., Lasky-Su, J., et al. (2008). Co-transmission of conduct problems with attention-deficit/hyperactivity disorder: familial evidence for a distinct disorder. Journal of Neural Transmission, 115(2), 163-175.

Collett, B. R., Ohan, J. L., & Myers, K. M. (2003). Ten-year review of rating scales. VI: scales assessing externalizing behaviors. Journal of the American Academy of Child and Adolescent Psychiatry, 42(10), 1143-1170.

Conners, C. K. (1997). Conners' Rating Scales-Revised: Technical Manual.

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

Conners, C. K., Sitarenios, G., Parker, J. D., & Epstein, J. N. (1998b). Revision and restandardization of the Conners Teacher Rating Scale (CTRS-R): factor structure, reliability, and criterion validity. Journal of Abnormal Child Psychology, 26(4), 279-291.

Eiraldi, R. B., Power, T. J., Karustis, J. L., & Goldstein, S. G. (2000). Assessing ADHD and comorbid disorders in children: the Child Behavior Checklist and the Devereux Scales of Mental Disorders. Journal of Clinical Child Psychology, 29, 3-16.

Gianarris, W. J., Golden, C. J., & Greene, L. (2001). The Conners' Parent Rating Scales: a critical review of the literature. Clinical Psychology Review, 21(7), 1061-1093.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. Journal of Child Psychology and Psychiatry, 38(5), 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. Journal of the American Academy of Child and Adolescent Psychiatry, 40(11), 1337-1345.

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(5), 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., 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(2), 129-134.

Greene, R. W., Biederman, J., Zerwas, S., Monuteaux, M. C., Goring, J. C., & Faraone, S. V. (2002). Psychiatric comorbidity, family dysfunction, and social impairment in referred youth with oppositional defiant disorder. American Journal of Psychiatry, 159(7), 1214-1224.

Hair, J. F., Black, W. C., Babin, B. E., Anderson, R. E., & Tatham, R. L. (2006). Multivariate Data Analysis (6th ed.). Upper Saddle River, N.J.: Prentice-Hall.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148, 839-843.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

Hudziak, J. J., Althoff, R. R., Stanger, C., van Beijsterveldt, C. E., Nelson, E. C., Hanna, G. L., et al. (2006). The Obsessive Compulsive Scale of the Child Behavior Checklist predicts obsessive-compulsive disorder: a receiver operating characteristic curve analysis. Journal of Child Psychology and Psychiatry, 47(2), 160-166.

Hudziak, J. J., Copeland, W., Stanger, C., & Wadsworth, M. (2004). Screening for DSM-IV externalizing disorders with the Child Behavior Checklist: a receiver-operating characteristic analysis. Journal of Child Psychology and Psychiatry, 45, 1299-1307.

Kraemer, H. C. (1992). Evaluating medical tests. Objective and quantitative guidelines. Newbury Park: Sage Publications, Inc.

Lahey, B. B., Loeber, R., Burke, J., Rathouz, P. J., & McBurnett, K. (2002). Waxing and waning in concert: dynamic comorbidity of conduct disorder with other disruptive and emotional problems over 7 years among clinic-referred boys. Journal of Abnormal Psychology, 111(4), 556-567.

Leon, A. C., Olfson, M., Weissman, M. M., Portera, L., & Sheehan, D. V. (1996). Evaluation of screens for mental disorders in primary care: methodological issues. Psychopharmacology Bulletin, 32(3), 353-361.

Marsh, H. W., Kit-Tai, H., & Zhonglin, W. (2004). In search of the golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings). Structural Equation Modeling, 11(3), 320-341.

Mathai, J., Anderson, P., & Bourne, A. (2004). Comparing psychiatric diagnoses generated by the Strengths and Difficulties Questionnaire with diagnoses made by clinicians. Australian and New Zealand Journal of Psychiatry, 38(8), 639-643.

Nock, M. K., Kazdin, A. E., Hiripi, E., & Kessler, R. C. (2007). Lifetime prevalence, correlates, and persistence of oppositional defiant disorder: results from the National Comorbidity Survey Replication. Journal of Child Psychology and Psychiatry, 48(7), 703-713.

Parker, J. D., Sitarenios, G., & Conners, C. K. (1996). Abbreviated Conners' Rating Scales revisited: A confirmatory factor analytic study. Journal of Attention Disorders, 1, 55-62.

Speltz, M. L., McClellan, J., DeKlyen, M., & Jones, K. (1999). Preschool boys with oppositional defiant disorder: clinical presentation and diagnostic change. Journal of the American Academy of Child and Adolescent Psychiatry, 38(7), 838-845.

Stringaris, A., & Goodman, R. (in press). Three dimensions of oppositionally in youth. Journal of Child Psychology and Psychiatry.

Taylor, E., Schachar, R., Thorley, G., & Wieselberg, M. (1986). Conduct disorder and hyperactivity: I. Separation of hyperactivity and antisocial conduct in British child psychiatric patients. British Journal of Psychiatry, 149, 760-767.

van der Oord, S., Prins, P. J., Oosterlaan, J., & Emmelkamp, P. M. (2006). The association between parenting stress, depressed mood and informant agreement in ADHD and ODD. Behaviour Research and Therapy, 44(11), 1585-1595.

van Lier, P. A., van der Ende, J., Koot, H. M., & Verhulst, F. C. (2007). Which better predicts conduct problems? The relationship of trajectories of conduct problems with ODD and ADHD symptoms from childhood into adolescence. Journal of Child Psychology and Psychiatry, 48(6), 601-608.

Waschbusch, D. A., Pelham, W. E., Jr., Jennings, J. R., Greiner, A. R., Tarter, R. E., & Moss, H. B. (2002). Reactive aggression in boys with disruptive behavior disorders: behavior, physiology, and affect. Journal of Abnormal Child Psychology, 30(6), 641-656.

7 General discussion

The present thesis deals with the role of self and parent rating scales in child and adolescent mental health assessments. For the identification of psychiatric disorders, standardized instruments like rating scales have become increasingly important in the light of evidence-based medicine. Clinical rating scales usually were tested for their underlying factor structure, for their internal consistency or for their associations with other rating scales. Thus, various reliability and validity measures were described that are helpful for the choice of the most accurate test. Furthermore, most rating scales have been analyzed in normative samples and T-scores for cut-off have been proposed in order to identify the most abnormal subjects. However, apart from T-scores and other validity measures the accordance of rating scale scores and diagnostic criteria for a psychiatric disorder is of particular interest in clinical settings.

The overarching goal of the present thesis was to test the diagnostic accuracy of clinical rating scales in the prediction of common child and adolescent psychiatric disorders. Knowledge of the diagnostic accuracy of clinical rating scales may improve an evidence-based psychiatric assessment for children and adolescents. For example, clinicians can decide for more accurate and more efficient rating scales. In consequence, time of the clinical assessments can be reduced and costs can be saved. In addition, problems to target in treatment can be identified faster and therapy can start earlier. Hence, the assessment for the child and adolescent and his family will be improved by minimizing time to suffer from the untreated disorder.

However, when testing the diagnostic accuracy of clinical rating scales some methodological and theoretical problems have to be solved as shown in chapter 3. Most notably, a diagnostic “gold standard” has to be defined. As mentioned in chapter 3.1, the validity of ICD-10 and DSM-IV diagnoses as a “gold standard” measure has been criticized and problems with structured and unstructured interviews for assessing diagnoses have been addressed.

Because this issue is threatening the results and conclusions of studies dealing with diagnostic accuracy it has to be discussed for the present findings.

In the present chapter the following issues have to be addressed: First, the general results and conclusions of the three previously described diagnostic accuracy studies will be summarized. Then, the actual limitations of the present studies will be discussed under the perspective of methodological considerations. Finally, a prospect of an improved psychiatric assessment in the light of evidence-based practice is given.

7.1 General conclusions of the present findings

7.1.1 Aims and methods of the studies

Study 1 and 3 aimed at testing diagnostic accuracy of three multivariate parent rating scales for the prediction of externalizing behavioral disorders. All of them, the Child Behavior Checklist (CBCL; Achenbach, 1991a), the Conners' Parent Rating Scale revised (CPRS-R; Conners, 1997) and the parent version of the Strength and Difficulties Questionnaire (PSDQ; Goodman, 1997, 1999) have been widely used in child and adolescent mental health clinics to evaluate diverse behavioral and emotional problems in youth.

In contrast to the first and the third studies, two self rating scales were tested in the second study: The Center of Epidemiological Studies Depression Scale (CES-D; Hautzinger & Bailer, 1993; Radloff, 1977) and the Youth Self Report (Achenbach, 1991b). For the prediction of internalizing disorders such as adolescent depression, information from the person concerned seems more accurate than parent or teacher information. This consideration was supported by previous findings (Berg-Nielsen, Vika, & Dahl, 2003; Zukauskiene, Pilkauskaite-Valickiene, Malinauskiene, & Krataviciene, 2004).

The CBCL (Achenbach, 1991a) as a parent rating scale and the corresponding YSR (Achenbach, 1991b) as a self rating scale have been tested in study 1 and 2. Both instruments have three levels of scoring: (1) eight primary scales named withdrawn, somatic, anxious/depressed, social problems, thought problems, attention problems, delinquent and aggressive behavior; (2) two second order scales called internalizing and externalizing and (3) a total problem score. In a recently presented revised version of the CBCL and the YSR (Achenbach, Dumenci, & Rescorla, 2003; Achenbach & Rescorla, 2001), new DSM-oriented scales have been introduced. Whereas the narrow-band scales were built on the basis of empirical data resulting from factor analyses, the DSM-oriented scales were built on expert opinion based on similarity to DSM-IV diagnostic criteria (Achenbach et al., 2003). As few studies tested these DSM-oriented scales before they were of particular interest for the present thesis.

The methods used in all three studies are comparable and have been described in chapter 3. For testing multivariate prediction models, logistic regression analyses were used for testing the accuracy of a scale. ROC analyses were used with the AUC as a measure of accuracy. Finally, cut-off analyses by the use of efficiency statistics were performed. Outside the focus of the present thesis, the third study has additionally addressed the construct validity of ODD.

7.1.2 General findings of the three studies

The results from study 1 and 2 convincingly show that the recently introduced DSM-oriented scales of the CBCL/YSR are much better suited for the prediction of ICD-10 and DSM-IV psychiatric disorders than the empirical scales. The validity of these DSM-oriented scales was partly confirmed and the tested scales can be recommended for clinical practice. The results from the third study confirmed the diagnostic accuracy of the PSDQ and the CPRS-R for assessing ODD in a sample of ADHD referred youth. Compared to the results of the second study, the diagnostic accuracy of the CPRS-R and the PSDQ is reduced but still sufficient. In addition, also the CBCL did not work as well as in the community sample when

testing it in a clinical sample. Apart from comorbid disorders which have been controlled in the present studies this may be due to the symptom overlap of psychiatric disorders in clinical samples. In total, the tested multidimensional rating scales, however, worked fine and were useful for diagnostic decision making. The advantages of multidimensional rating scales and nosological implications of the present findings are presented below.

7.1.3 Preference for multidimensional rating scales

In conclusion, the present thesis can confirm the diagnostic validity of multidimensional rating scales (MRS) addressing diverse emotional and behavioral problems. All of the three tested MRS (e. g. CBCL, YSR, CPRS-R and PSDQ) include specific scales that were found accurate for the prediction of psychiatric disorders according to ICD-10 or DSM-IV criteria. Multidimensional instruments are advantageous compared to one-dimensional rating scales which are focused exclusively on a specific disorder. First, MRS are more practicable in clinical institutions as they are easier to handle than diverse separate scales. Secondly, most psychiatric disorders in children and adolescents are not occurring segregated and are accompanied by further emotional and behavioral symptoms or comorbid psychiatric disorders (e.g. Angold, Costello, & Erkanli, 1999). Thus, MRS can provide more detailed information about the youths' mental health. Thirdly, one dimensional rating scale may be more biased by halo and priming effects due to the greater transparency of the scales to the tested person. In contrast, the items of a scale in MRS can be presented shuffled in combination with items from other scales. Hence, the validity of a scale can be improved. For more detailed information on the construction of rating scales, the interested reader is referred to the specific literature (e.g. Spector, 1992).

7.1.4 Nosological implications

Reduced but still acceptable diagnostic accuracy was found in samples with patients suffering from various mental disorders. Various somatic diseases may be identified more

accurately due to more precise symptom definitions and aetiologally defined diagnostic entities. However, in psychiatry,the causes of emotional and behavioral disorders often remain less clear than the causes of disorders in somatic medicine and some patients show symptoms of multiple disorders. Furthermore, diagnostic criteria often include items that are not specific to a certain diagnosis. For example, both ICD-10 and DSM-IV consider the presence of concentration problems in ADHD and in clinical depression. In addition, irritability may be part of mania, clinical depression and ODD according to ICD-10 and DSM-IV. However, the type of concentration problems or the type of irritability can differ between disorders. For example, different forms of irritability have been described (e.g. chronic vs. episodic; Leibenluft, Cohen, Gorrindo, Brook, & Pine, 2006). Thus, further studies providing explanation about the time frame and the detailed phenomenology of symptoms are necessary. For the development of the upcoming classification systems, i.e. ICD-11 and DSM-V (Kupfer, Regier, & Kuhl, 2008), criteria on symptoms should be described more precisely for a better nosological understanding of psychiatric disorders. Tests of construct validity as used in the third study of the present thesis may be helpful to identify dimensions of disorders and to gain more detailed knowledge of psychopathology.

7.2 Limitations of the present studies

7.2.1 The problem of the "gold standard"

As mentioned in the introduction chapter (see 3.1), the "gold standard" is of general importance in studies dealing with diagnostic accuracy. In psychiatry, the "gold standard" has been defined by ICD-10 or DSM-IV criteria and this information has come from an adolescent or a parent diagnostic interview. In the first and the third study, a diagnostic interview with a parent was used. In the second study, a case-control design with probands from community sample and patients referred for clinical depression was used. This heterogeneous sample may be criticized for not being accurate to the sample the scale should be used later (Bossuyt et al., 2003; Gray, 2004). In addition, the second study was based on expert

clinicians' diagnoses with questionable reliability. Nevertheless, for an exploratory test of the scales' diagnostic performance, this procedure may be useful (Sullivan Pepe, 2003).

In addition, also the diagnostic assessment by structured interviews has been criticized regarding the issues of validity (Gray, 2004). Thus, information from structured interviews may be biased by the informant and interviewer or the questions for diagnostic criteria remain unclear for the informant and provide unspecific information.

7.2.2 The problem of information sources

Until now, the question which informants should be considered and how inconsistent information should be assessed remains unsolved. In practice, a multi-informant assessment of childhood problems is complicated since ratings from informants and from different settings are often only modestly associated (Kerr, Lunkenheimer, & Olson, 2007; Offord et al., 1996). At one time, this disagreement was thought to reflect the unreliability of measures and informant bias. However, discrepancies can also reflect the true variation in children's' behaviors across diverse settings and relational circumstances (Stanger, Achenbach, & McConaughy, 1993). Taking these aspects into account, the results of the present findings in all three studies are limited: either only one informant has been considered or clinical expert diagnoses with unclear reliability have been used. However, the clinical expert diagnoses were based on multivariate information sources as a parent, child and teacher information. In addition, behavioral observations from the clinician were considered for diagnostic decision making. The added value of these clinical diagnoses was limited by less evident combination of this information in order to come to a final decision. However, the missing information algorithm may affect the reliability of these clinical diagnoses.

Overall, a standard for multi-informant diagnostic decision making in child and adolescent psychiatry is missing. Therefore, structured interviews for both child and parent and

additional diagnostic questions for teachers and caregivers are necessary. In addition, a diagnostic algorithm has to be defined in order to come to a general diagnostic decision.

7.3 Implications for improved evidence-based assessments

Overall, the results of the three previous described studies did not suggest the use of the rating scales as a surrogate of a comprehensive clinical assessment. However, the scale with most adequate diagnostic accuracy was recommended for clinical practice in combination with further tests and interviews. Apart from the score of the rating scale, further information about the onset, duration and impairment is necessary to identify the final psychiatric disorder. For example, the symptoms of separation anxiety disorders as difficulty in separating at night or an inappropriate fear of being alone can be elevated but these problems may have started after the child's 7th birthday. Therefore, no ICD-10 disorder of separation anxiety can be given due to the defined criteria. Furthermore, other strains or a traumatic experience of the child may cause symptoms similar to early childhood separation anxiety. The discrimination of early separation anxiety and a trauma-caused disorder may be important because different treatment procedures will be necessary (e.g. parent training vs. trauma therapy).

Therefore, improved assessments include multidimensional and multi-informant rating scales (MMRS) with clearly defined symptoms oriented towards ICD and/or DSM criteria. For conflicting information between different informants, either advice for the most relevant informant should be presented or a weighted score should be applied. For both of these possibilities, further research is needed. However, the CBCL and the YSR included in the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001) the CPRS-R included in the Conners' Rating Scales (CRS; Conners, 1997) and the PSDQ as the parent version of the Strength and Difficulties Questionnaire (SDQ; Conners, 1997; Goodman, 1997, 1999) represent such MMRS. They all provide information about the most common psychiatric disorders but with the exception of the SDQ no algorithm has been

proposed in order to come to a final psychiatric diagnosis. Therefore, according to the suggestions above the SDQ seems to be the most promising MMRS. However, this thesis did not test the multivariate and multi-informant SDQ diagnoses algorithms in the prediction of multiple psychiatric diagnoses.

In addition to the previous suggestions, two possible implications for an improved assessment are proposed. First, MMRS should be used as initial screening instruments , cut-off scores have to be defined in order to maximize sensitivity, and additional testing has to be made in order to come to a final diagnosis. Secondly, the MMSR should be completed by additional information about the onset, duration and impairment of the symptoms. Furthermore, a defined computer algorithm for making a diagnosis should be applied.

7.3.1 MMRS as initial screening instruments in psychiatric assessments

If MMRS are used as initial screening inventories in child and adolescent psychiatric assessments, it is important to identify all possible subjects with the disorder and, thus, a high sensitivity of these instruments is needed. As shown in chapter 3.3.2, both sensitivity and specificity have to be considered when evaluating a clinical rating scale. However, when looking for the most accurate cut-off score sensitivity has to be weighted higher on the costs of reduced specificity. Thus, cut-off scores by quality efficiency statistics are not accurate because specificity and sensitivity have been considered equally according to the base rate of the disorder in the sample. For the solution of this problem, a weighted efficiency indicator can be applied as described by Kraemer (1992). However, a weighted efficiency indicator requires information about the costs and benefits of false positive and false negative classified patients and frequently these figures are not available. As an alternative, we have proposed to define a range of acceptable cut-off scores as shown in the second study. The lower bound of the defined range is representing the most efficient and optimized sensitive measure for the diagnosis that has to be predicted. Finally, it is possible to use stratum specific likelihood ratios instead of specific cut-off scores. Likelihood ratios have been

introduced in chapter 3.3.3 and the statistics for the calculation of the optimal ranges has been shown by Radack, Rouan & Hedges (1986). Thus, when using a MMRS as a screening instrument a lower likelihood range can be defined to identify most people at risk for the disorder.

A positive test result of the MMRS can be defined as a starting point of a further clinical assessment. In consequence, further information on age of onset, continuity, impairment, specificity of symptoms and information about other psychiatric disorders should be included in order to arrive at the final diagnosis.

7.3.2 Advanced and comprehensive MMRS as diagnostic tools

More recent MMRS not only include diagnosis-oriented scales but also further information about the onset, duration and impact of the disorder. Furthermore, they define a mathematical algorithm how this information was included for a comprehensive diagnostic decision. Some of these instruments are already available.

a) Strength and Difficulties Questionnaire

First, as shown before the SDQ is such an instrument. The diagnosis algorithm in order to come to probable or possible psychiatric diagnosis including information about the duration and impact of the symptoms is presented in the appendix. Further information is presented on http://www.sdqinfo.com. First studies show that the SDQ system is an accurate and very useful instrument for psychiatric decision making in children and adolescents by using the proposed computer algorithm and multiple informants (Goodman, 1999; Goodman, Ford, Simmons, Gatward, & Meltzer, 2003; Goodman, Renfrew, & Mullick, 2000; Mathai, Anderson, & Bourne, 2004). In the third study of this thesis, the computer algorithm has been partly confirmed by the parent version. However, the studies confirming the SDQ are limited by some methodological constraints. For example, the New Zealand Study (Mathai et al., 2004) is not reporting specificity values so that the results are difficult to interpret.

Furthermore, the SDQ is limited because only diagnostic categories (e.g. affective disorders, attention disorders, behavioral disorders) can be predicted. In addition and because the scales are short and easy to use, only 5 items have been used in order to predict the diagnostic category. Although, the SDQ does include further information about the onset, duration and impact of the symptoms this instrument is suggested as a screening instrument rather than a comprehensive diagnostic tool (Goodman et al., 2003).

b) Development and Well-Being Assessment

The Development and Well-Being Assessment (DAWBA; Goodman, Ford, Richards, Gatward, & Meltzer, 2000) is a comprehensive diagnostic instrument for children and adolescents which has been recently introduced. Teacher, parent and adolescent information is considered in order to come to a psychiatric diagnosis according to ICD-10 or DSM-IV. Furthermore, it includes the SDQ, the family background of the youth and information about the strength and resources of the adolescent. The DAWBA interviews can be administered either by humans or by computers. The internet form of the DAWBA (see http://www.dawba.com) is easy to handle and includes a mixture of closed questions such as "Does he ever worry?" and open-ended questions such as "Please describe in your own words what it is that he worries about?". Information from different informants is compiled by a computer program that also predicts the likely diagnosis or diagnoses. Afterwards, experienced clinical raters decide whether to accept or overturn the computer diagnoses (or lack of diagnoses) in the light of their review of the full data including the free text passages. The DAWBA was used in epidemiological and clinical samples and was found as a useful instrument in both settings (Goodman, Ford et al., 2000). However, studies concerning the validation of the diagnosis are limited to date.

7.4 Implication for future studies concerning diagnostic accuracy

Further studies of diagnostic accuracy are needed because of the increased importance of rating scales in psychiatric assessments. However, the “gold standard” of the future studies

should be based on multi-informant diagnoses especially when testing MMRS. Furthermore, the validity of ICD and DSM diagnoses should be improved and more detailed criteria on the presence of symptoms should be made available. If the “gold standard” can be improved, more valid information on diagnostic accuracy of a rating scale is possible. The recently introduced DAWBA is an instrument that may set new standards for psychiatric diagnoses in children and adolescents.

Furthermore, the standards of diagnostic accuracy studies as described by Bossuyt et al. (2003) should be considered and have to be adapted to studies testing diagnostic accuracy of psychiatric disorders.

7.5 References

Achenbach, T. M. (1991a). Manual for the Child Behavior Check List/4-18 and 1991 Profile. Burlington, VT: Department of Psychiatry, University of Vermont.

Achenbach, T. M. (1991b). Manual for the Youth Self Report and 1991 Profile. Burlington, VT: Department of Psychiatry, University of Vermont.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). DSM-oriented and empirically based approaches to constructing scales from the same item pools. J Clin Child Adolesc Psychol, 32(3), 328-340.

Achenbach, T. M., & Rescorla, L. A. (2001). Manual for the School-Age Forms and Profiles. Child Behavior Checklist. Teacher's Report Form. Youth Self-Report. An Integrated System of Multi-informant Assessment. Burlington: Library of Congress.

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. Journal of Child Psychology and Psychiatry, 40(1), 57-87.

Berg-Nielsen, T. S., Vika, A., & Dahl, A. A. (2003). When adolescents disagree with their mothers: CBCL-YSR discrepancies related to maternal depression and adolescent self-esteem. Child Care Health Dev, 29(3), 207-213.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Bmj, 326(7379), 41-44.

Conners, C. K. (1997). Conners' Rating Scales-Revised: Technical Manual.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. Journal of Child Psychology and Psychiatry, 38(5), 581-586.

Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. J Child Psychol Psychiatry, 40(5), 791-799.

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(5), 645-655.

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2003). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Int Rev Psychiatry, 15(1-2), 166-172.

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(2), 129-134.

Gray, G. E. (2004). Evidence Based Psychiatry. Washington DC: American Psychiatric Publishing.

Hautzinger, M., & Bailer, M. (1993). Allgemeine Depressions-Skala (ADS). Deutsche Form der " Center of Epidmiologic Studies Depression Scale" (CES-D). Weinheim: Beltz.

Kerr, D. C., Lunkenheimer, E. S., & Olson, S. L. (2007). Assessment of child problem behaviors by multiple informants: a longitudinal study from preschool to school entry. J Child Psychol Psychiatry, 48(10), 967-975.

Knotterus, J. A. (2002). The Evidence Base of Clinical Diagnosis. London: BMJ Books.

Kraemer, H. C. (1992). Evaluating medical tests. Objective and quantitative guidelines. Newbury Park: Sage Publications, Inc.

Kupfer, D. J., Regier, D. A., & Kuhl, E. A. (2008). On the road to DSM-V and ICD-11. Eur Arch Psychiatry Clin Neurosci, 258 Suppl 5, 2-6.

Leibenluft, E., Cohen, P., Gorrindo, T., Brook, J. S., & Pine, D. S. (2006). Chronic versus episodic irritability in youth: a community-based, longitudinal study of clinical and diagnostic associations. J Child Adolesc Psychopharmacol, 16(4), 456-466.

Mathai, J., Anderson, P., & Bourne, A. (2004). Comparing psychiatric diagnoses generated by the Strengths and Difficulties Questionnaire with diagnoses made by clinicians. Australian and New Zealand Journal of Psychiatry, 38(8), 639-643.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., et al. (1996). Integrating assessment data from multiple informants. J Am Acad Child Adolesc Psychiatry, 35(8), 1078-1085.

Radack, K. L., Rouan, G., & Hedges, J. (1986). The likelihood ratio. An improved measure for reporting and evaluating diagnostic test results. Arch Pathol Lab Med, 110(8), 689-693.

Radloff, L. S. (1977). The CES-D scale: a self report depression scale for research in general populations. Appl. Psychol. Meas., 1, 385-401.

Spector, P. (1992). Summated rating scale construction. Newbury Park: Sage Publications.

Stanger, C., Achenbach, T. M., & McConaughy, S. H. (1993). Three-year course of behavioral/emotional problems in a national sample of 4- to 16-year-olds: 3. Predictors of signs of disturbance. J Consult Clin Psychol, 61(5), 839-848.

Sullivan Pepe, M. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford: University Press.

Zukauskiene, R., Pilkauskaite-Valickiene, R., Malinauskiene, O., & Krataviciene, R. (2004). Evaluating behavioral and emotional problems with the Child Behavior Checklist and Youth Self-Report scales: cross-informant and longitudinal associations. Medicina (Kaunas), 40(2), 169-177.

8 Curriculum vitae

Name	AEBI, Marcel
Date of birth	August 22, 1971
Place of birth	Zurich
Place of citizenship	Heimiswil (Be), Zurich
Nationality	Swiss

1988-1993	Gymnasium, Zurich (Matura Type C)
1994-2001	Psychology studies at the University of Zurich with a major in Developmental Psychology 1st minor: Psychopathology of Children and Adolescents 2nd minor: Criminology
2000-2001	Management support in the Department of Human Resources at the Credit-Suisse Financial Services
2001-2009	Clinical Psychologist, Child and Adolescent Psychiatric Service of the Canton of Zurich
2002- present	Postgraduate Education in Behavioral Therapy for Children and Adolescents, Academy of Behaviour Therapy in Children and Adolescents, Universities of Zurich, Basel and Fribourg
2005-present	Research Assistant and Assistant Lecturer, Division of Child and Adolescent Psychopathology, Department of Child and Adolescent Psychiatry, University of Zurich
2009-present	Research Coordinator, Division of Forensic Psychiatry, Department of Child and Adolescent Psychiatry, University of Zurich

9 Appendix

9.1 STARD checklist for reporting diagnostic accuracy studies

Section and topic	Item	Description
Title, abstract, and keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity")
Introduction	2	State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups
Methods:		
Participants	3	Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected
	4	Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Describe participant sampling: was this a consecutive series of participants defined by selection criteria in items 3 and 4? If not, specify how participants were further selected
	6	Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	Describe the reference standard and its rationale
	8	Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both
	9	Describe definition of and rationale for the units, cut-off points, or categories of the results of the index tests and the reference standard
	10	Describe the number, training, and expertise of the persons executing and reading the index tests and the reference standard
	11	Were the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to the readers.
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (eg 95% confidence intervals)
	13	Describe methods for calculating test reproducibility, if done
Results:		
Participants	14	Report when study was done, including beginning and ending dates of recruitment
	15	Report clinical and demographic characteristics (eg age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centre)
	16	Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test results	17	Report time interval from index tests to reference standard, and any treatment administered between
	18	Report distribution of severity of disease (define criteria) in those with the target condition and other diagnoses in participants without the target condition
	19	Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, report the distribution of the test results by the results of the reference standard
	20	Report any adverse events from performing the index test or the reference standard
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)
	22	Report how indeterminate results, missing responses, and outliers of index tests were handled
	23	Report estimates of variability of diagnostic accuracy between readers, centres, or subgroups of participants, if done
	24	Report estimates of test reproducibility, if done
Discussion	25	Discuss the clinical applicability of the study findings

9.2 Computer algorithm in SPSS for scoring the SDQ

The scoring algorithm is based on the 25 variables plus impact items for each questionnaire. The algorithm expects to find these variables with specific names: the first letter of each variable name is 'p' for the parent SDQ, 's' for the self-report SDQ and 't' for the teacher SDQ. After this first letter, the variable names are as follows:

consid = Item 1 : considerate
restles = Item 2 : restless
somatic = Item 3 : somatic symptoms
shares = Item 4 : shares readily
tantrum = Item 5 : tempers
loner = Item 6 : solitary
obeys = Item 7 : obedient
worries = Item 8 : worries
caring = Item 9 : helpful if someone hurt
fidgety = Item 10 : fidgety
friend = Item 11 : has good friend
fights = Item 12 : fights or bullies
unhappy = Item 13 : unhappy
popular = Item 14 : generally liked
distrac = Item 15 : easily distracted
clingy = Item 16 : nervous in new situations
kind = Item 17 : kind to younger children
lies = Item 18 : lies or cheats
bullied = Item 19 : picked on or bullied
helpout = Item 20 : often volunteers
reflect = Item 21 : thinks before acting
steals = Item 22 : steals
oldbest = Item 23 : better with adults than with children
afraid = Item 24 : many fears
attends = Item 25 : good attention
ebddiff = Impact question: oveall difficulties in at least one area
distres = Impact question: upset or distressed
imphome = Impact question: interferes with home life
impfrie = Impact question: interferes with friendships
impclas = Impact question: interferes with learning
impleis = Impact question: interferes with leisure

For each of these items, if the first response category (not true, no, not at all) has been selected, this is coded as zero, the next response category (somewhat true, yes-minor, just a little) is coded as one and so on.
For each informant, the algorithm generates six scores. The first letter of each derived variable is 'p' for parent-based scores, 's' for self-report-based scores and 't' for teacher-based scores. After this first letter, the names of the scores are as follows:

emotion = emotional symptoms
conduct = conduct problems
hyper = hyperactivity/inattention
peer = peer problems
prosoc = Prosocial
ebdtot = total difficulties
impact = Impact

*** Recoding variables and then scoring the parent SDQ scores

```
SET FORMAT=F8.0.
RECODE pobeys (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO qobeys .
EXECUTE .
RECODE pfriend (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO qfriend .
EXECUTE .
RECODE ppopular (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO qpopular .
EXECUTE .
RECODE preflect (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO qreflect .
EXECUTE .
RECODE pattends (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO qattends .
EXECUTE .
RECODE pdistres (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO qdistres .
EXECUTE .
RECODE pimphome (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO qimphome .
EXECUTE .
RECODE pimpfrie (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO qimpfrie .
EXECUTE .
RECODE pimpclas (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO qimpclas .
EXECUTE .
RECODE pimpleis (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO qimpleis .
EXECUTE .

COMPUTE pemotion = RND(MEAN.3(psomatic,pworries,punhappy,pclingy,pafraid) * 5) .
EXECUTE .
COMPUTE pconduct = RND(MEAN.3(ptantrum,qobeys,pfights,plies,psteals) * 5) .
EXECUTE .
COMPUTE phyper = RND(MEAN.3(prestles,pfidgety,pdistrac,qreflect,qattends) * 5) .
EXECUTE .
COMPUTE ppeer = RND(MEAN.3(ploner,qfriend,qpopular,pbullied,poldbest) * 5) .
EXECUTE .
COMPUTE pprosoc = RND(MEAN.3(pconsid,pshares,pcaring,pkind,phelpout) * 5) .
EXECUTE .
COMPUTE pebdtot = SUM.4(pemotion,pconduct,phyper,ppeer) .
EXECUTE .
COMPUTE pimpact = SUM.1(qdistres,qimphome,qimpfrie,qimpclas,qimpleis) .
EXECUTE .
IF (pebddiff=0) pimpact=0 .
EXECUTE .
DELETE VARIABLES qobeys qreflect qattends qfriend qpopular qdistres qimphome qimpfrie qimpclas
qimpleis .
```

```
*** Recoding variables and then scoring the self-report SDQ scores

SET FORMAT=F8.0.
RECODE sobeys (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO robeys .
EXECUTE .
RECODE sfriend (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO rfriend .
EXECUTE .
RECODE spopular (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO rpopular .
EXECUTE .
RECODE sreflect (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO rreflect .
EXECUTE .
RECODE sattends (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO rattends .
EXECUTE .
RECODE sdistres (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO rdistres .
EXECUTE .
RECODE simphome (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO rimphome .
EXECUTE .
RECODE simpfrie (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO rimpfrie .
EXECUTE .
RECODE simpclas (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO rimpclas .
EXECUTE .
RECODE simpleis (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO rimpleis .
EXECUTE .

COMPUTE semotion = RND(MEAN.3(ssomatic,sworries,sunhappy,sclingy,safraid) * 5) .
EXECUTE .
COMPUTE sconduct = RND(MEAN.3(stantrum,robeys,sfights,slies,ssteals) * 5) .
EXECUTE .
COMPUTE shyper = RND(MEAN.3(srestles,sfidgety,sdistrac,rreflect,rattends) * 5) .
EXECUTE .
COMPUTE speer = RND(MEAN.3(sloner,rfriend,rpopular,sbullied,soldbest) * 5) .
EXECUTE .
COMPUTE sprosoc = RND(MEAN.3(sconsid,sshares,scaring,skind,shelpout) * 5) .
EXECUTE .
COMPUTE sebdtot = SUM.4(semotion,sconduct,shyper,speer) .
EXECUTE .
COMPUTE simpact = SUM.1(rdistres,rimphome,rimpfrie,rimpclas,rimpleis) .
EXECUTE .
IF (sebddiff=0) simpact=0 .
EXECUTE .
DELETE VARIABLES robeys rreflect rattends rfriend rpopular rdistres rimphome rimpfrie rimpclas
rimpleis .

*** Recoding variables and then scoring the teacher SDQ scores

SET FORMAT=F8.0.
RECODE tobeys (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO uobeys .
EXECUTE .
RECODE tfriend (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO ufriend .
EXECUTE .
RECODE tpopular (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO upopular .
EXECUTE .
RECODE treflect (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO ureflect .
EXECUTE .
RECODE tattends (0=2) (1=1) (2=0) (ELSE=SYSMIS) INTO uattends .
EXECUTE .
RECODE tdistres (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO udistres .
EXECUTE .
RECODE timpfrie (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO uimpfrie .
EXECUTE .
RECODE timpclas (0=0) (1=0) (2=1) (3=2) (ELSE=SYSMIS) INTO uimpclas .
EXECUTE .
```

```
COMPUTE temotion = RND(MEAN.3(tsomatic,tworries,tunhappy,tclingy,tafraid) * 5) .
EXECUTE .
COMPUTE tconduct = RND(MEAN.3(ttantrum,uobeys,tfights,tlies,tsteals) * 5) .
EXECUTE .
COMPUTE thyper = RND(MEAN.3(trestles,tfidgety,tdistrac,ureflect,uattends) * 5) .
EXECUTE .
COMPUTE tpeer = RND(MEAN.3(tloner,ufriend,upopular,tbullied,toldbest) * 5) .
EXECUTE .
COMPUTE tprosoc = RND(MEAN.3(tconsid,tshares,tcaring,tkind,thelpout) * 5) .
EXECUTE .
COMPUTE tebdtot = SUM.4(temotion,tconduct,thyper,tpeer) .
EXECUTE .
COMPUTE timpact = SUM.1(udistres,uimpfrie,uimpclas) .
EXECUTE .
IF (tebddiff=0) timpact=0 .
EXECUTE .
DELETE VARIABLES uobeys ureflect uattends ufriend upopular udistres uimpfrie uimpclas .
```