In the absence of a gold standard

Noordhof, Arjen

Drukker : Boxpress BV - Oisterwijk

Vormgever & omslag : Wytse Noordhof - Dieren

ISBN : 978-90-367-4180-4 (druk)
978-90-367-4179-8 (digitaal)

RIJKSUNIVERSITEIT GRONINGEN

In the Absence of a Gold Standard

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
woensdag 3 februari 2010
om 13.15 uur

door

Arjen Noordhof
geboren op 12 april 1980
te Rheden

Promotores :	Prof. dr. J. Ormel Prof. dr. A. J. Oldehinkel
Beoordelingscommissie :	Prof. dr. W. A. M. Vollebergh Prof. dr. R. R. Meijer Prof. dr. H. M. Koot

How pathetically scanty my self-knowledge is compared with, say, my knowledge of my room. There is no such thing as observation of the inner world, as there is of the outer world.

- Franz Kafka

The research presented in this thesis is part of the TRacking Adolescents' Individual Lives Survey (TRAILS). Participating centers of TRAILS include various departments of the University Medical Center and University of Groningen, the Erasmus University Medical Center Rotterdam, the University of Utrecht, the Radboud Medical Center Nijmegen, and the Trimbos Institute, all in the Netherlands. Principal investigators are prof. dr. J. Ormel (University Medical Center Groningen) and prof. dr. F.C. Verhulst (Erasmus University Medical Center). TRAILS has been financially supported by various grants from the Netherlands Organization for Scientific Research NWO (Medical Research Council program grant GB-MW 940-38-011; ZonMW Brainpower grant 100-001-004; ZonMw Risk Behavior and Dependence grants 60-60600-98-018 and 60-60600-97-118; ZonMw Culture and Health grant 261-98-710; Social Sciences Council medium-sized investment grants GB-MaGW 480-01-006 and GB-MaGW 480-07-001; Social Sciences Council project grants GB-MaGW 457-03-018, GB-MaGW 452-04-314, an GB-MaGW 452-06-004; NWO large-sized investment grant 175.010.2003.005); the Sophia Foundation for Medical Research (projects 301 and 393), the Dutch Ministry of Justice (WODC), and the participating universities. I am grateful to all adolescents, their parents and teachers who participated in this research and to everyone who worked on this project and made it possible.

I want to thank the following institutions for contributing to the printing costs of this thesis: The graduate school for Behavioural and Cognitive Neurosciences (BCN), the University Medical Center Groningen (UMCG), and the Rijksuniversiteit Groningen (RuG).

Contents

Chapter 1

Introduction

There is no gold standard for measurement and diagnosis of psychopathology. An example of a gold standard is an answer sheet to a math exam. If the answers given by a respondent differ from the sheet, the respondent made mistakes. The reason we can treat this as a gold standard is that we have a strong believe in the correctness of mathematical proof. In psychiatry and psychopathology there is no such answer sheet and therefore no quick and easy way to ascertain whether the results of a test or a diagnosis are right. If discrepancies arise between different diagnostic procedures there is no final judgment to which both results can be compared to decide which one is true.

This does not imply that all measures are equally important, useful or valid, nor that the phenomena captured under the umbrella of psychopathology are not disturbing. I assume that the concept psychopathology refers to real problems that can be meaningfully communicated between people and influence the results obtained from measurement instruments. That being said, we are far away from a clear and certain understanding of what it is that we are measuring and communicating. In the papers presented in this thesis I aimed to confront a few of these disturbing uncertainties: (1) uncertainty about the latent structures underlying the covariances between measures of psychopathology, (2) uncertainty about the estimation of these covariances, (3) discrepancies between informants, and (4) the diagnostic concepts used in clinical practice. Admittedly, these four topics cannot be dealt with in full depth in a single thesis and the papers may elicit more questions than answers. However, I hope the thesis also shows the interrelatedness of these issues and their importance for applied psychological science. In the following I will shortly introduce each of these four topics.

Latent structures underlying psychopathology

Mental disorders cannot be observed, but diagnostic concepts are derived from ideas about (causal) associations between observations. Therefore diagnoses are sometimes referred to as hypothetical constructs (Strauss & Smith, 2009). Hypothetical emphasizes that we hypothesize that the terms refer to real attributes (Borsboom, Mellenbergh, & van Heerden, 2004), while the term construct refers to the fact that we make them ourselves while developing our sciences and practices. Some constructs are defined as dichotomies, which implies the hypothesis that the attribute is either present or not. Some constructs are defined as ordinal or continuous. In that case the attribute is hypothesized to be present in all people, albeit in different quantities. Finally, we can imagine mixtures in which a construct may only apply to subsets of people, but occur in different quantities within these people.

Psychological constructs are often developed on the basis of covariances between reported symptoms in a sample from the general population. The basic idea underlying this approach is that differences between individuals (inter-individual differences) result from the same underlying causal system. Differences in observed variables are assumed to be caused by differences in latent variables related to these underlying causal systems. For example, Watson, Wiese, Vaidya, & Tellegen (1999) hypothesize that observed differences in reported emotions are caused by differences in the functioning of two biological systems (Gray, 1990): the behavioral activation system (BAS) and behavior inhibition system (BIS). Individual differences in the functioning of these two systems would explain why persistently a two-dimensional structure of affect is found in factor-analytic studies of differences in self-reported emotion. To be sure, the above mentioned factor-analyses are only part of the argument for BIS and BAS. Latent factors derived from statistical analysis cannot be directly interpreted as indicators of existing phenomena, but are themselves in need of explanation.

With regard to psychopathology, the currently used Diagnostic Statistical Manual (4th ed.; American Psychiatric Association, 1994) provides diagnostic rules for research and clinical practice that allow users to apply diagnostic constructs to individuals. The validity of these constructs has been a continuous topic of debate among experts. Particularly, some have argued that there is no strong support for the dichotomous nature of many of the constructs and that the boundaries between constructs have not been shown to 'carve nature at its joints' (Lilienfeld, Waldman, & Israel, 1994; Waller, 2006). From a practical point of view it has been observed that many people meet criteria for multiple diagnoses, which is generally referred to as comorbidity. Given the problematic status of the validity of the dichotomous constructs, this comorbidity should not necessarily be interpreted as the presence of two diseases within the same individual (see Neale & Kendler, 1995). As an alternative some authors have used factor-analysis to study the latent structure underlying the DSM-IV comorbidity patterns (e.g. Vollebergh, et al., 2001). Also, authors have developed questionnaires independent of the DSM-IV system and used factor-analysis to investigate the structure of covariance between items on these questionnaires (e.g. Achenbach, 1991a; Achenbach & Edelbrock, 1978; Hartman, et al., 1999). Many DSM-IV and questionnaire based studies have resulted in a similar two-dimensional second-order latent variable model that has been found for children and adolescents (e.g. Lahey, et al., 2008), as well as adults (Krueger, Caspi, Moffitt, & Silva, 1998; Krueger, Chentsova-Dutton, Markon, Goldberg, & Ormel, 2003). The term 'second-order' refers to the fact that these two dimensions capture the covariance-structure of first-order subscales, which capture the covariance between symptoms (i.e. cognitions, emotions, behaviors). This model is generally referred to as the structure of Internalizing (INT) and Externalizing (EXT) psychopathology. In my view the single most important advantage of latent variable models is that they make a formal distinction between latent variables and observed questionnaire responses. This also implies that observed data can be used to compare alternative latent models, as will be done in chapter 2.

Second, they capture multiple psychological concepts within the same analytic model. Different psychological concepts (e.g. temperament and personality) have often been treated in separate literatures in which the overlap with other constructs (e.g. mental disorders) was neglected (Clark, 2005). This process can easily result in a multitude of ill-understood concepts that cause more confusion than clarification of the latent structure involved. Latent variable modeling of the covariance structure of multiple constructs of psychopathology can be used to investigate both common and specific features of these concepts. For example, the concept of a Broader Autism Phenotype (BAP) has received much attention in the literature on autism and has been conceptualized as a specific dimension or trait in the general population. In chapter 2 it is shown that the problems related to this dimension can be adequately studied within the framework of Internalizing and Externalizing psychopathology. This type of analysis is useful in order to evaluate how the BAP-concept relates to other concepts of psychopathology. A third advantage is that the variables in these models have been constructed as dimensions. As will be argued in chapter 3 there are important advantages to first develop a dimensional representation and only later introduce categories within this dimensional framework.

Discrepancies in the estimation of association between measures

Latent structures of psychopathology, like the model of internalizing and externalizing psychopathology, are generally derived from the covariance structure of reported symptoms in a sample. This means that it is assumed that the covariances are caused by individual differences in underlying psychopathology. However, estimations of associations between variables may also be influenced by several methodological biases. Specifically, sampling and measurement biases may influence the estimated association between different reported emotional and behavioral problems. In chapter 5 it will be tested to what extent these biases influence the estimated association between the internalizing and externalizing dimensions of psychopathology.

Discrepancies between informants

Symptoms of psychopathology are not directly observed by researchers or diagnosticians. Furthermore, one of the key criteria of DSM-IV for the application of any diagnosis is that the syndrome "… causes clinically significant distress or impairment in social, occupational, or other important areas of functioning" . That is, we need to know not only what behaviors and emotions occur in daily life, but also their impact on the person and the environment. The researcher is therefore dependent on informants, i.e. either self-report or reports of people related to the subject. Either through interviews or questionnaires informants are asked to report on the behaviors and emotions of the subject. This inevitably introduces informant-specific sources of variance in questionnaire responses. The often low correlations (e.g. r=.30; Achenbach, McConaughy, & Howell, 1987) between different informant reports suggest powerful informant-specific influences. If different informant reports would

reliably and validly measure one and the same attribute one would not expect such low correlations. For this reason researchers have attempted to distinguish between variance that is caused by characteristics of the subject and variance that is caused by the specific informants being used. In this thesis several of these models are applied and evaluated (chapter 4, 5, and 6).

It is important to realize that informant discrepancies do not necessarily indicate influences of methodological factors. In chapters 4 to 6 multiple reasons will be discussed why discrepancies may arise between informants. Of specific importance in chapter 4 are differences in the context in which an informant observes the subject and differences in the perspective by others and the self-perspective. These differences are not well captured by the term bias as they refer to actual observations rather than misrepresentation of observations. This does not mean that biases in observing and in responding to questionnaires are absent. In the chapters of this thesis both substantive and methodological reasons for the emergence of discrepancies will be discussed.

The creation of useful diagnostic language for clinical practice

Given the uncertain status of the latent models and hypothetical constructs of psychopathology one may wonder whether and how psychological problems can be meaningfully communicated among experts and between experts and non-experts. The important critiques that have accumulated over the years with regard to the dichotomies of DSM-IV suggest that this model may give a misleading impression of knowledge about psychiatric disorders. The manual itself makes it explicit that "in DSM-IV, there is no assumption that each category of mental disorder is a completely discrete entity with absolute boundaries dividing it from other disorders or from no mental disorder." (American Psychiatric Association, 1994) Nevertheless, the language that is developed by strictly applying the diagnostic rules is one of discrete disorders and 'comorbidity'. Furthermore, in many countries practices have developed that attach much status to DSM-IV dichotomous diagnoses which can influence treatment and reimbursement. The implication is that a gap may grow between the apparently solid categorical diagnostic language and the dimensional and uncertain knowledge about the latent structure to which this language should refer.

There is no quick and easy way to bridge this gap. As already mentioned, there is no gold standard to which we can refer for absolute measures and there is much uncertainty about the validity of diagnoses made in clinical practice. In communicating with clients in clinical practice it is nearly impossible to discuss all the subtle and less subtle arguments for and against certain diagnostic approaches. Nevertheless, I think it is crucial that a more nuanced and realistic diagnostic language be developed. In chapter 3 a perspective is developed on how to create useful diagnostic language on the basis of a dimensional framework of psychopathology. In the conclusion to this thesis I discuss whether and how uncertainties about measurement and discrepancies between informants can be incorporated into this diagnostic language.

Chapter 2

Integrating the 'Broader Autism Phenotype' into General Dimensional Frameworks of Psychopathology.

Arjen Noordhof, Robert F. Krueger, Johan Ormel, Albertine J. Oldehinkel, Catharina A. Hartman *

Abstract

The concept of a 'Broader Autism Phenotype', a dimensional approach to problems related to Autism Spectrum Disorders (ASD), has received much attention in recent literature. ASD-problems occur frequently in the general population and often co-occur with problems from other domains of psychopathology. In the research presented here these co-occurrence patterns were investigated by integrating a dimensional approach to ASDs into more general dimensional frameworks of psychopathology. Factor Analysis was used to develop models covering multiple domains of psychopathology. A bi-factor model of specific and non-specific features of psychopathology showed the most adequate model fit in three measurement waves of a longitudinal general population sample (N=2230, ages 10-17). The results show that (a) problems traditionally related to the domain of ASDs can be adequately integrated into general population based dimensional models of psychopathology, (b) the 'Broad Autism Phenotype' can be regarded as a specific domain of problems that can be distinguished from the domains of Internalizing, Externalizing and Attention Problems, and (c) specific subdomains of BAP are differently related to INT, EXT and Attention Problems.

Introduction

Autism Spectrum Disorders (ASDs) are characterized by problems from the domains of (1) reciprocal social behavior, (2) language development and communication, or (3) repetitive/stereotypic behavior. Some children with these problems can be diagnosed according to the narrow criteria of 'Autistic Disorder', but a larger group of children show problems from these domains that do not meet these narrow criteria. This observation has led to the idea of a spectrum (Wing & Gould, 1979) which consists of both narrowly defined autistic disorder and milder forms of autistic problems. In DSM-IV some (but not all) children with these milder problems can be diagnosed as suffering from 'Asperger's disorder' or 'Pervasive Developmental Disorder- Not otherwise Specified' (PDD-NOS). PDD-NOS constitutes a residual, catch-all category that does not correspond to a clearly defined disorder and a

* *Information about all co-authors of the articles in this thesis can be found on page 106*

dimensional conceptualization is a promising alternative in capturing the structure of these problems without imposing arbitrary cut-offs. A dimensional approach also accommodates continuity between PDD-NOS, subthreshold symptoms, and normality (Constantino & Todd, 2003). Furthermore, dimensions are in line with the finding that the family members of children with a diagnosis of autism often present with milder symptoms of what has been called the 'Broader Autism Phenotype' (BAP; Bolton, et al., 1994; Folstein & Rutter, 1988).

The research we report here is aimed at testing and expanding the dimensional approach by integrating autism spectrum problems into more general dimensional frameworks of psychopathology. Dimensional models of psychopathology are particularly useful in that they can provide insight into the high comorbidity rates that have been found between ASDs and other domains of psychopathology (de Bruin, Ferdinand, Meester, de Nijs, & Verheij, 2007; Simonoff, et al., 2008). Furthermore, they illuminate the dimensional structure of autism spectrum-problems by testing the hypothesis that they constitute a single coherent spectrum of problems in the general population.

Comorbidity among syndromes is not specific to ASDs, but a general phenomenon in psychiatric classification (Kessler, Chiu, Demler, & Walters, 2005). If syndromes have clearly distinct underlying causal factors, comorbidity indicates the presence of two diseases at the same time. In psychiatry this is seldom the case and comorbidity may indicate that the syndrome categories used do not adequately capture the underlying causal structures from which problems emerge (Krueger & Markon, 2006a; Meehl, 2001). For understanding comorbidity in psychiatry it is therefore useful to reconsider and remodel the way psychopathology is conceptualized. Research on general population data shows that many cross-diagnostic correlation patterns can be captured by a structure with two higher-order-factors, generally referred to as the Internalizing (INT) and Externalizing (EXT) spectrum (Krueger, et al., 1998; Krueger, et al., 2003; Vollebergh, et al., 2001). This work builds on work by Achenbach and colleagues who were the first to propose this distinction between Internalizing and Externalizing problems in child psychiatry (Achenbach, 1966). Recently, a similar model has been found to fit data on child- and adolescent DSM-IV syndromes (Lahey, et al., 2008).

Comorbidity in these higher-order models may be interpreted in many other ways than the idea of 'two diseases at the same time' and has resulted in new theories about the causal structure of psychopathology (e.g. Krueger & Markon, 2006b). Comparable, but slightly different from the higher order factor models are models of specific and non-specific features of psychopathology (Weiss, Susser, & Catron, 1998). In these models comorbidity is interpreted as indicating that part of the problems are non-specific, i.e. not indicating a syndrome-specific latent structure, and therefore their variance is primarily related to non-specific variation in the amount of problems. Variance not captured by this component is interpreted as indicating a more specific latent structure related to broad domains like INT or EXT, or specific domains like

'Depression' or 'Aggression'. These models have been applied by using contrasts (Essex, et al., 2006; Weiss, et al., 1998) or Principal Component Analysis (Noordhof, Oldehinkel, Verhulst, & Ormel, chapter 4 of this thesis). For the research reported here we used Confirmatory Factor Analysis (CFA), because it allows for a direct comparison with the aforementioned higher-order models (Patrick, Hicks, Nichol, & Krueger, 2007; Yung, Thissen, & McLeod, 1999).

ASDs show high comorbidity with other DSM-IV syndromes (de Bruin, et al., 2007; Simonoff, et al., 2008). Furthermore, behavioral-genetic studies have shown substantial shared additive genetic variance of the BAP and other domains of psychopathology (Hoekstra, Bartels, Hudziak, van Beijsterveldt, & Boomsma, 2007; Reiersen, Constantino, Grimmer, Martin, & Todd, 2008; Ronald, Simonoff, Kuntsi, Asherson, & Plomin, 2008). Thus, BAP problems are both phenotypically and genotypically related to other domains of psychopathology. Integration of these problems into the general INT and EXT dimensional models of psychopathology described above may offer new insights into the phenotypic structure. A second rationale for the integration of BAP into existing dimensional models is to understand whether the higher order structure of BAP problems is appropriately conceptualized as a single dimension. Some authors have found that the first principal component captured a major part of variance in ASD-problems (Constantino, et al., 2004). Others, using other instruments, have found solutions with three or more correlated factors (e.g. Boomsma, et al., 2008; Hoekstra, Bartels, Cath, & Boomsma, 2008; Luteijn, Luteijn, Jackson, Volkmar, & Minderaa, 2000; Volkmar, et al., 1988). Happe and Ronald (2008) have argued that the correlations between the different subdomains within the BAP are actually quite low and suggestive of a 'fractionable autism triad'. This triad is proposed to consist of relatively independent subdomains which are related to different specific underlying (biological) mechanisms. In a recent review Mandy and Skuse (2008) concluded that only a few studies have directly addressed the hypothesis that the social (i.e. reciprocal behavior, social information processing) and non-social (i.e. restricted interest, repetitive behavior) problems are strongly related and that most of the evidence does not support a strong link between these domains. Nevertheless, the authors conclude that completely abandoning the idea of a 'Broad Autism Phenotype' is premature, given the inconclusiveness of the evidence and the fact that the loose covariance between sub domains is still a finding that deserves research attention. Some authors have proposed a higher-order model (Bolton, et al., 1994) in which BAP subscales load on one higher-order factor. However, it is not certain that a BAP factor will actually emerge as a distinct factor, in joint analysis with subscales from other domains of psychopathology. Therefore, the present joint analysis of BAP, Internalizing, and Externalizing subscales in the general population can also be used to further test the hypothesis that the multiple subdomains that are considered part of the BAP constitute a distinct and coherent dimension which can be differentiated from INT and EXT.

To summarize, dimensional conceptualization of problems that are traditionally referred to as the 'Broader Autism Phenotype' offers a promising approach to increase

understanding of the latent structure of these problems in the general population. Integration into more general dimensional frameworks may offer insight into the structure of psychopathology in general and the structure of ASD-problems in particular. To this end, we developed and tested several factor analytic models in a general population cohort of (pre)adolescents. On the one hand these analyses can be regarded as an exploration of the covariance structure underlying these multiple problem domains. On the other hand these models allowed us to test the specific hypothesis that INT, EXT and BAP constitute three, correlated, problem domains in the general population, or alternatively, that covariance between problems from these domains is either suggestive of a more simple structure (e.g., BAP problems can be fully captured by INT and EXT factors) or a more complex structure. (e.g., BAP problems cannot be captured by a single higher order factor).

Methods

Sample

Subjects were participants in the 'Tracking Adolescents' Individual Lives Survey' (TRAILS), a prospective multi-cohort study of Dutch (pre)adolescents. The study involved a representative sample from the general population and is described in detail in Huisman et al.(2008). Briefly, the target sample involved all 10- to 11-year-old children living in the three largest cities and some rural areas in the North of The Netherlands. Of the eligible children, 76.0% (n=2230, mean age = 11.09, SD =0.55) were enrolled in the study. Responders and non-responders did not differ regarding the prevalence of teacher-rated problem behavior and associations between sociodemographic variables and mental health indicators (De Winter, et al., 2005). To date, the population cohort has been assessed three times (T1: March 2000- July 2001, T2: September 2003- December 2004, T3: September 2005-December 2007). Participation rates were 96.4% at T2 (mean age= 13.55, SD = 0.53), and 81.4% at T3 (mean age= 16.25, SD = 0.73). After complete description of the study to the subjects, written informed consent was obtained from the parents at each assessment wave and from the adolescents at T2 and T3. T1, T2, and T3 data are used in the present study.

Instruments

- *CBCL*

The Dutch version of the Child Behavior Checklist (CBCL; Achenbach, 1991a; Verhulst, van der Ende, & Koot, 1996) was used to assess Internalizing, Externalizing and Attention problems. The CBCL is a 112-item questionnaire on which parents rate descriptions of emotions and behaviors on a 3-point scale (not [0], sometimes [1], or very often [2]). The period over which they are asked to report is the last six months. In the TRAILS-study the questionnaire was completed by one of the parents, which was the mother in most cases. Factor analysis on these items revealed a structure of eight syndrome scales (Achenbach, 1991a). Three of the CBCL scales are related to the

Internalizing domain (INT): Anxious-Depressed (13 items, α=0.78), Somatic complaints (11 items, α=0.69), and Withdrawn-Depressed (8 items, α=0.71). Two are related to the Externalizing domain (EXT): Aggressive Behavior (18 items, α=0.88) and Rule-Breaking behavior (17 items, α=0.68). The other three scales are Attention Problems (10 items, α=0.82), Social Problems (11 items, α=0.78), and Thought Problems (15 items, α=0.63). In a study by Hartman et al. (1999) the distinction between an INT and EXT factor was replicated quite well, although they found no significant difference in model fit between a 2-factor and an 8-factor solution.

The scales Thought Problems and Social Problems are generally not regarded part of either the INT or EXT domain. Integration of these diverse items into INT, EXT or more comprehensive dimensional models may be interesting for its own sake, but was regarded too complex within the context of the current paper. The Attention Problems scale is also not part of INT or EXT in the CBCL. However, this scale is related to ADHD, which is part of the EXT-spectrum in other studies (e.g. Lahey, et al., 2008) and which is strongly comorbid with ASDs (Simonoff, et al., 2008). Therefore the Attention Problems scale was included in the analysis.

• *CSBQ*

The parent-rated Child Social Behavior Questionnaire (Luteijn, Jackson, Volkmar, & Minderaa, 1998) was used to assess problems that are commonly found in children diagnosed with an ASD. The instrument has a 3-point rating-scale that is equal to the CBCL-format (not [0], sometimes [1], or very often [2]). The CSBQ differs from the CBCL in that the questions refer to a two-month period. Originally, the instrument consisted of 96 items covering the full range of problems, with an emphasis on the milder variants seen in PDD-NOS (Luteijn, et al., 1998). Hartman, Luteijn, Serra, & Minderaa (2006) refined and shortened the CSBQ to 49 items and found a 6-factor structure with Exploratory Factor Analysis (EFA): Behavior/Emotions not Optimally Tuned to the Social Situation (Not tuned, 11 items, α=0.84), Reduced Contact and Social Interests (Reduced Contact, 12 items, α=0.76), Orientation Problems in Time, Place or Activity (Orientation, 8 items, α=0.78), Difficulties in Understanding Social Information (Social Understanding, 7 items, α=0.75), Stereotyped Behavior (Stereotyped, 8 items, α=0.69), and Fear and Resistance to Change (Fear of Change, 3 items, α=0.74).

Statistical analyses

MPlus version 5.2 was used to explore and test latent variable models of the subscales of the CBCL and the CSBQ. All scales were skewed and some scales were extremely skewed (skewness coefficient > 2): CBCL Rule-Breaking, and CSBQ Orientation, Social Understanding, Stereotyped, and Fear of Change. To accommodate for this, we transformed all variables by taking their natural logarithms.[1]

[1] *We added 1 to the scores before taking the natural logarithm, because the natural logarithm of zero is undefined.*

Furthermore, we used maximum likelihood estimation with robust standard errors (MLR), because of its relative robustness to deviations from normality. We used the Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI) as indicators of absolute model fit. The Bayesian Information Criterion (BIC) was used to compare different models with satisfactory RMSEA and CFI. We started by developing models on T2-data and then replicated these model for the younger (T1) and older (T3) measurement waves.

First, a higher-order model was developed on the basis of T2-data. To this end we used both CFA and EFA. We started with testing the higher-order model illustrated in Figure 1, which corresponds to the hypothesis that INT, EXT and BAP are three different, correlated, domains of psychopathology. Subsequently, we explored and tested multiple alternative models, which will be described in more detail in the results-section. Finally, we selected those models that showed adequate model fit for replication in T1 and T3 data.

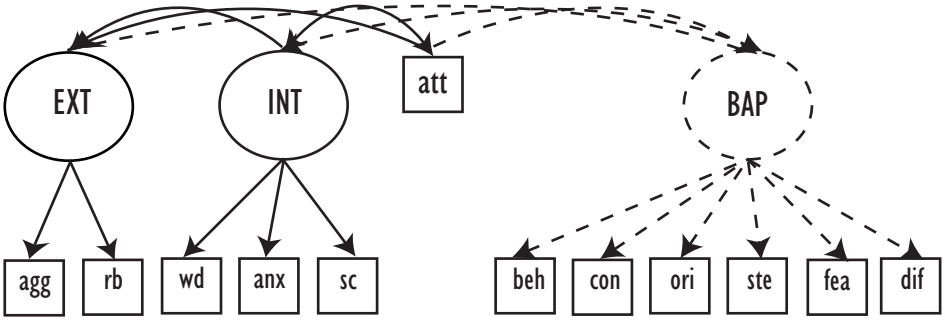

Figure 1. Higher order model with and without inclusion of a factor related to the 'Broader Autism Phenotype' (indicated by dotted lines).
Note: EXT = Externalizing,; *INT* = Internalizing; *BAP* = Broader Autism Phenotype; *Agg* = Aggressive Behavior; *Anx* =Anxious-Depressed; *Rb* = Rule-Breaking Behavior; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed; *Att* = Attention Problems; *Beh* = Behavior and Emotions Not Optimally Tuned to The Social Situation; *Dif* = Difficulties in Understanding Social Information; *Fea* = Fear of and Resistance to Changes; *Ori* = Orientation-problems in time, place, or activity; *Con* = Reduced Contact and Social Interests; *Ste* = Stereotyped behavior.

Second, a bi-factor model was developed. A bi-factor model corresponds to the hypothesis that part of the variance in problem behavior consists of non-specific (NS) covariance between all subscales. Covariance that is not captured by this NS-factor is hypothesized to be related to a specific problem domain. In contrast with higher-order models, bi-factor model factors are orthogonal (i.e. the factors are not correlated). Instead, the correlations between the factors in the higher-order model are assumed to be captured by the NS-factor in the bi-factor model. We tested multiple alternative

bi-factor models and selected those that showed adequate model fit for replication in T1 and T3 data.

Third, we evaluated whether the selected models could be replicated in different measurement-waves within the same sample. We compared fit indices of the multiple alternative models in T1, T2, and T3 in order to test whether the same model was superior in all measurement waves.

Results

Developing a higher-order model on T2-data

We developed higher-order models by fitting CFA-models on the one hand and exploring possible improvements using EFA on the other. This approach allowed us to test specific hypotheses regarding BAP and its co-occurrence with other problem domains and develop a well-fitting integrative higher-order model. The analysis proceeded in five steps and resulted in the selection of two adequate models to be replicated on T1 and T3 data. All fit indices of these analyses are shown in Table 1.

In step one, we tested the basic hypothesis that the CSBQ-scales measure a single distinct problem domain (BAP) and that comorbidity of BAP with Internalizing and Externalizing problems can be captured by correlations with the INT and EXT higher-order factors. To this end the 'basic higher-order model' (Figure 1) was tested with the use of CFA. In this model Attention Problems was regarded as a different problem domain indicated by only one subscale and correlated with the three higher order factors. An adequate fit of this model would support the strategy of simply investigating INT, EXT and BAP as three distinct domains of psychopathology measured by two different instruments. However, as shown in Table 1, this model did not fit well to the T2-data (RMSEA>.05; CFI<.95). This implies that the covariance between CSBQ- and CBCL-scales cannot be regarded as simply reflecting correlations between broad domains of psychopathology and should be investigated in more detail, which was done in step two and three.

In step two, we explored the relation between subscales in more detail, using EFA to model the covariance structure of CSBQ and CBCL-scales. A well-fitting 5-factor model (RMSEA<.05; CFI>.95) was found with a lower BIC-value than the 4-factor model (see Table 1). The 6-factor model did not converge, so we used the factor loadings of the 5-factor model. Factor loadings of the 5-factor model are shown in Table 2.

Table 1. *Fit indices for CFA and EFA-models developed on T2-data and replicated on T1 and T3 data.*

Data	Model	Step[a]	Figure[b]	RMSEA	CFI	BIC
T2	Higher-order	1	1	0.12	0.86	37886
	EFA					
	4 factor	2		.047		41656
	5 factor			.03		41588
	6 factor			nc		
	Higher-order					
	Refined	3	2	0.05	0.98	36710
	Without BAP	4	2	0.12	0.86	37899
	Third-order	5	3	0.06	0.98	36730
	Bi-factor					
	Refined	1	4	0.05	0.98	36655
	Without BAP* [c]	2	4	0.08	0.95	36943
	Without F4*	3	5	0.06	0.98	36726
T1	Higher order					
	Refined		2	0.05	0.98	41954
	Third-order		3	0.05	0.97	41963
	Bi-factor					
	Refined		4	0.04	0.99	41863
	Without F4*		5	0.05	0.98	41937
T3	Higher-order					
	Refined		2	0.06	0.97	29127
	Third-order		3	0.06	0.97	29142
	Bi-factor					
	Refined		4	0.05	0.99	29042
	Without F4*		5	0.05	0.98	29099
	Without F4*		5	0.05	0.98	29099

Note: *RMSEA* = Root Mean Square Error of Approximation; *CFI* = Comparative Fit Index; *BIC* = Bayesian Information Criterion; *EFA* =Exploratory Factor Analysis; *BAP* = Broader Autism Phenotype; *F4* = Factor from previous EFA.

[a] Refers to the analytic steps described in the results-section.

[b] Refers to the Figure in which the model is shown.

[c] A * refers to factors in the bi-factor models which correspond to, but are not equivalent to, the factors with the same names in the higher-order models.

Table 2. *Loadings of subscales in 5-factor Exploratory Factor Analysis-solution.*

	F1	F2	F3	F4	F5
Agg		0.88			
Anx	0.80				
Rb		0.53		0.31	
Sc	0.48				
Wd	0.38		0.69		
Att				0.64	
Beh		0.71			0.34
Dif				0.32	0.40
Fea	0.34				0.42
Ori				0.58	0.42
Con			0.46		0.46
Ste				0.21	0.44

Note: Only loadings >.2 are reported. Factor Fx refers to the x-th factor derived in the EFA-solution; *Agg* = Aggressive Behavior; *Anx* =Anxious-Depressed; *Rb* = Rule-Breaking Behavior; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed; *Att* = Attention Problems; *Beh* = Behavior and Emotions Not Optimally Tuned to The Social Situation; *Dif* = Difficulties in Understanding Social Information; *Fea* = Fear of and Resistance to Changes; *Ori* = Orientation-problems in time, place, or activity; *Con* = Reduced Contact and Social Interests; *Ste* = Stereotyped behavior.

In step three, we constructed a 'refined higher-order model' in order to model the more specific relations between CBCL- and CSBQ-scales that were suggested by EFA. A first refinement was to include loadings of CSBQ-scales on the higher-order factors INT and EXT. Based on the EFA-factors F1 and F2, free loadings were added of the CSBQ-scale Fear of Change on INT and Not Tuned on EXT. This indicates that in the refined model these two CSBQ-scales are interpreted as partly related to the INT and EXT domain rather than only related to the BAP domain. As a second refinement, a free correlation was added between the residual variances of the CSBQ-scale Reduced Contact and that of the CBCL-scale Withdrawn-Depressed, because the EFA-factor F3 suggested a specific covariance between these two scales.

The third refinement was to add the fourth EFA-factor to the model by specifying a higher-order factor F4 with loadings of five subscales. The EFA-factor F4 captures covariance between two CBCL (Rule-Breaking Behavior, and Attention Problems) and three CSBQ-scales (Orientation, Stereotyped, and Social Understanding). These three additions resulted in the 'refined higher-order model' illustrated in Figure 2. As shown in Table 1, this model fitted well to the T2-data (RMSEA=.05, CFI>.95) and had a lower BIC-value than the basic model. This means that the EFA-based improvements were sufficient to adequately capture the covariance between CBCL and CSBQ in a higher-order model.

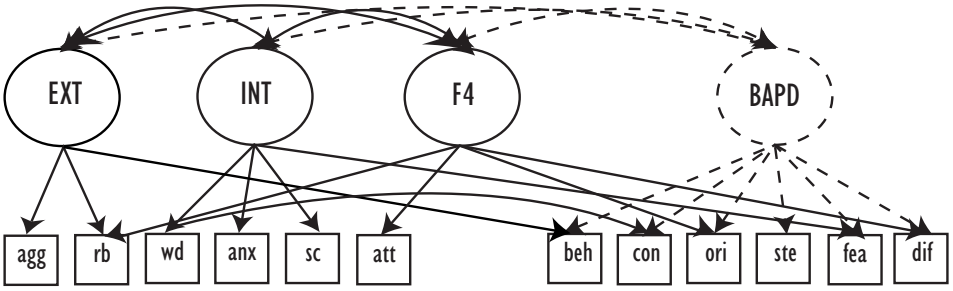

Figure 2. Refinement of the higher order model on the basis of interpretation of EFA.
Note: Abbreviations are explained in figure 1.

In step four, we tested the hypothesis that the six CSBQ-scales had to be modeled as a separate domain in the general population. For this aim we tested whether a model without a BAP-factor fitted the data. This was not the case (Table 1: RMSEA>.05; CFI<.95) and the BIC-values were even lower than those for the basic model. This shows that it was necessary to include a BAP-factor, which implies that the six CSBQ subscales constitute a specific domain that can be distinguished from INT and EXT.

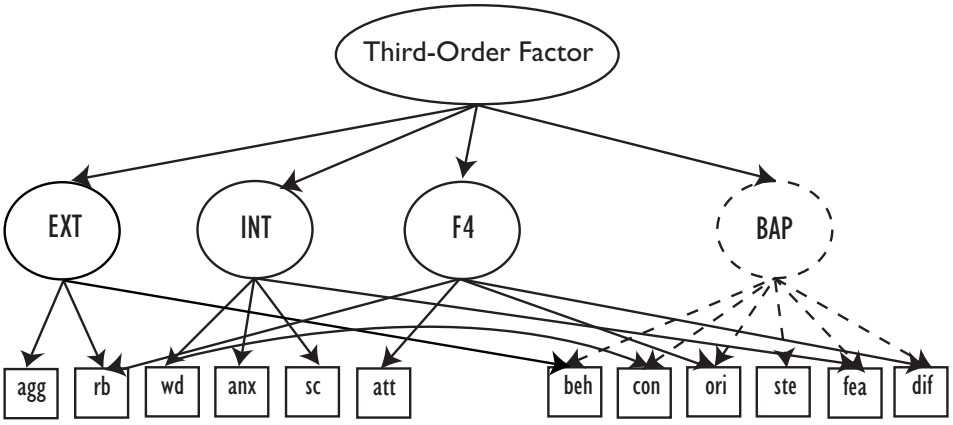

Figure 3. The refined higher order model with inclusion of a third-order factor.
Note: Abbreviations are explained in figure 1; *F4* = Factor from previous Exploratory Factor Analysis.

In step five, we tested the hypothesis that the covariance between the four higher-order domains (INT, EXT, BAP and F4) could be captured by a single third-order factor (see Figure 3). As described in the introduction section of this paper, this model can be conceptually and statistically distinguished from the way between-domain covariance is captured in a bi-factor model (see next paragraph). Fitting both models to T2-data allowed for a direct comparison between their fit indices. Including a third-order factor resulted in slightly worse fit indices (see Table 1). This shows that covariance between the four higher-order factors could not be fully captured by a single third-order factor.

On the basis of these five analyses, the 'refined higher-order model' with inclusion of a BAP-factor (step 3) was chosen as the most adequate model and selected for replication in T1 and T3-data. The model with addition of the third-order general factor (step 5) was also selected for replication, because the fit indices of this model were only slightly worse.

Developing a bi-factor model on T2-data

The bi-factor model was developed as an alternative to the higher-order models. This was done to compare two different interpretations of the covariance between INT, EXT and BAP-problems. In the higher-order models, covariance between these domains is interpreted as a correlation between higher-order factors (Figure 2) or as an expression of an underlying, more general, third-order factor (Figure 3). In a bi-factor model the correlation between domain-specific factors is fixed at zero and a non-specific factor is introduced on which all subscales may have loadings. Therefore, between-domain covariance is interpreted as 'non-specific' covariance between subdomains rather then as correlation between higher-order factors. We developed a bi-factor model in three steps.

In the first step, the 'refined higher-order model' (Figure 2) was 'translated' into a bi-factor model to allow for a direct comparison of model fit between these two models. To this end we simply added a NS-factor and fixed all correlations between factors at 0 (see Figure 4).[2] This model showed adequate model fit (RMSEA=.05, CFI=.98).

In the second step, we tested the effect of removing the BAP*-factor from the model. Removal of the BAP-factor had resulted in completely inadequate model fit for the higher-order model (see step 4 above). However, this does not automatically generalize to the bi-factor model, because the NS-factor may capture part of the covariance between the six CSBQ subscales, which is captured by the BAP-factor in the higher-order model. A bi-factor model without BAP* did not show adequate fit indices (RMSEA= .08, CFI= .95), which supports the presence of covariance between all six CSBQ-scales that is not captured by the non-specific covariance between all CBCL and CSBQ-scales.

[2] *Because of introducing the NS-factor and orthogonality, the factors in the bi-factor are not the same as those in the higher-order model. To distinguish between the two a * is added to the factors of the bi-factor model.*

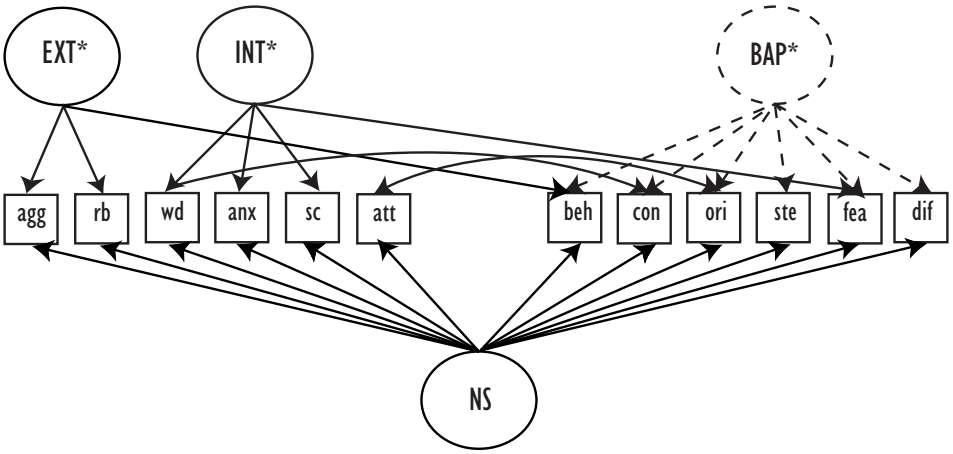

Figure 4. The bi-factor model based on adding an NS-factor to the 'refined higher-order model' and fixing correlations between factors at zero.
Note: Abbreviations are explained in figure 1. A * refers to a factor in the bi-factor that corresponds to, but is not equivalent to, the factor with the same name in the higher-order models; *F4* = Factor from previous Exploratory Factor Analysis.

The third step was based on an exploratory finding from observing the factor loadings of the bi-factor model with an additional F4* factor (Figure 4). It was found that the scales Social Understanding and Stereotyped had very low loadings (<.15) on this F4* factor. The loading of the Rule-Breaking scale was also rather low (.22). On the basis of this observation, we developed an alternative model, which includes a correlation between the Attention Problems and the Orientation scale rather than an F4*-factor. The model is illustrated in Figure 5. Fit indices for this model were only slightly inferior (RMSEA=.06, CFI=.98).

On the basis of these three analyses the 'refined bi-factor model' (Figure 4) was selected for replication, because the fit indices for this model were best. Furthermore, the model with only a correlation between Attention and Orientation Problems (Figure 5), rather than an F4* factor, was selected as it showed only slightly worse fit indices.

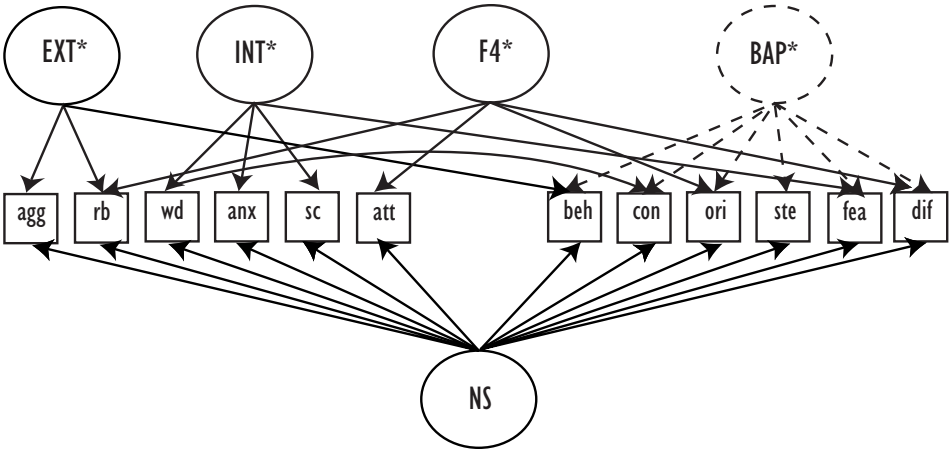

Figure 5. Alternative bi-factor model in which factor F4* is replaced by a free correlation between Att and Ori.
Note: Abbreviations are explained in figure 1.

Fitting models on T1 and T3 data

After having selected two higher-order models (Figure 2 and 3) and two bi-factor models (Figure 4 and 5), we tested the same models in different measurement waves of the sample. This was done to test whether the findings replicated in a younger (T1) and older (T3) age group, and to compare the models and test whether the same model showed superior fit indices at T1, T2 and T3.

As shown in table 1, CFI fit indices were adequate for all models (CFI>.95). Only the refined bi-factor model with F4* showed completely satisfactory RMSEA indices (RMSEA<.05). Furthermore, for this model BIC-values were lowest at all occasions. These results show that the bi-factor model shown in figure 4 is preferable to the other models in terms of model fit, but the differences shown in Table 1 are only modest. Factor loadings and residual variances of this model in the T2-data are shown in table 3.[3] This model includes the three specific factors INT*, EXT*, and BAP*. Furthermore, it includes two exploratory findings. First, a free correlation between the Withdrawn-Depressed and the Reduced Contact subscales. Second, the factor F4*, which is dominated by the subscales Attention Problems and Orientation. These two exploratory findings will be interpreted in the discussion.

[3] *In view of the limited space for this article we do not report the loading-patterns for T1 and T3 data. These were very comparable, but not exactly equal. These results can be obtained from the authors upon request.*

Table 3. *Factor loadings of the final bi-factor model for T2-data and explained variance for each subscale.*

	NS		EXT* [a]		INT*		F4*		BAP*		R^2
Agg	1.00	*0.77*	1.00	*0.55*							0.89
Anx	0.71	*0.63*			1.00	*0.56*					0.71
Rb	0.67	*0.64*	0.52	*0.35*			1.15	*0.22*			0.58
Sc	0.41	*0.40*			0.47	*0.35*					0.25
Wd	0.60	*0.59*			0.56	*0.29*					0.47
Att	0.41	*0.68*					1.00	*0.33*			0.57
Beh	0.88	*0.70*	0.74	*0.42*					1.00	*0.26*	0.74
Dif	0.68	*0.65*					0.71	*0.13*	1.01	*0.31*	0.53
Fea	0.30	*0.45*			0.33	*0.31*			0.76	*0.37*	0.43
Ori	0.58	*0.58*					2.40	*0.48*	1.27	*0.41*	0.73
Con	0.63	*0.55*							1.39	*0.39*	0.46
Ste	0.36	*0.45*					0.53	*0.13*	0.88	*0.35*	0.34

Residual Correlation Wd with Con = *0.56* [b]

Note: Standardized loadings are shown in italics; *NS* = Non Specific; *EXT** = Externalizing; *INT** = Internalizing; *F4** = Factor modeled on the basis of the fourth factor in the earlier Exploratory Factor Analysis; *BAP** = Broader Autism Phenotype; Other abbreviations are explained in table 2.

[a] A * refers to factors in the bi-factor models which correspond to, but are not equivalent to, the factors with the same names in the higher-order models.

[b] Residual correlation between Wd and Con was freely estimated.

Discussion

This study shows that ASD problems can be adequately integrated into dimensional frameworks of psychopathology. Support was found for the concept of a 'Broader Autism Phenotype' (BAP; Bolton, et al., 1994; Folstein & Rutter, 1988) in the general population that can be distinguished from the Internalizing and Externalizing spectrum. We also found support for the importance of subscale specific processes, as emphasized in the idea of a 'Fractionable Autism Triad' (Happe & Ronald, 2008). More generally, the study provides insight into the co-occurrences of BAP and its subdomains with other domains of psychopathology. Before discussing these insights, we will highlight some limitations of the study and discuss some precautions in interpreting the results.

Limitations

This is the first study that fully integrated BAP into higher-order and bi-factor models of Internalizing and Externalizing problems. The generalizability of the results may therefore be limited by the specifics of our study and replications are necessary before reaching more definite conclusions. These specifics include sample (Northern

part of the Netherlands), age (11-17), measures (CSBQ and CBCL), and informant (parent).

A second limitation is that the models are based on subscales rather than items. The drawback is that covariance between subscales of the CSBQ and CBCL may in some cases result from overlapping item content. Such item overlap is suggested by an EFA by Hartman et al. (2006). A full-scale bottom-up analysis of the two instruments could be used to resolve this issue. However, that is a computationally and conceptually complex project, which was beyond the scope of the current paper.

Finally, the research is limited by the fact that we only used factor analysis. Therefore, all the models we compared assumed normally distributed latent variables with linear relations to the observed variables. There may be underlying phenomena in our data that are better captured by non-linear relations or categorical variables. In future studies categorical latent variables may be studied by employing the techniques of latent class analysis or factor mixture modeling (Lubke & Muthen, 2005).

Interpreting the bi-factor model

The best fitting model was a bi-factor model. Before interpreting the specific factors in this model one may wonder what the 'non-specific' factor (NS) is. First, analyses of subscales that are all positively correlated will generally result in a first unrotated factor on which all subscales load substantially, which is the primary reason to call it non-specific. Second, systematic biases and specific viewpoints of the informant may result in increased correlations between subscales in non-specific ways and therefore contribute to the variance of the NS-factor. Third, by face-value inspection of the items and intuitive thinking one can think of many different ways in which these problems can be causally related in the development of a child and how multiple biological and social systems may interact in this development. A manifold of small positive causal effects crossing the borders of specific domains of psychopathology can result in the positive correlations that are found between subscales and therefore result in a large NS-factor (van der Maas, et al., 2006).

Our findings are much in line with the conclusion drawn in a recent review by Mandy and Skuse (2008) that, although subscale specific processes may be of primary importance, abandoning the concept of a 'Broader Autism Phenotype' is premature and neglects the modest coherence that is found between subdomains. The finding of a BAP-factor independent of NS, INT and EXT supports this conclusion. Yet the results also show that BAP cannot be understood well if it is studied as a single trait in the general population (Happe & Ronald, 2008; Mandy & Skuse, 2008). Subscales of the CSBQ are differentially related to other domains of psychopathology, and may involve different traits in the general population rather than just one single problem domain. In what follows, the relation of BAP-subdomains with the Internalizing, Externalizing and Attention Problems domains and implications for understanding comorbidity will be discussed.

BAP and the Internalizing Domain

Children with ASD often meet criteria for an anxiety disorder, specifically Social Anxiety Disorder (de Bruin, et al., 2007; Simonoff, et al., 2008). These children seldom meet criteria for a depressive disorder (de Bruin, et al., 2007; Simonoff, et al., 2008), but others have reported higher rates of co-occurrence between depression and ASDs (see Ghaziuddin, Ghaziuddin, & Greden, 2002). Brereton, Tonge, and Einfeld (2006) showed that prevalence of depressive symptoms increased with age, so the finding of a low co-occurrence may be specific for children. The other way around, ASD-symptoms have been found to be substantially elevated in children with diagnoses of Internalizing disorders and most strongly in mood disorders (Pine, Guyer, Goldwin, Towbin, & Leibenluft, 2008; Towbin, Pradella, Gorrindo, Pine, & Leibenluft, 2005). The current study shows two specific factors that may improve understanding of comorbidity.

First, the scale Reduced Contact was specifically correlated with Withdrawn-Depressed, but did not load on the INT-factor. Reduced Contact contains several items specifically related to core features of ASDs (e.g. 'makes little eye contact', 'dislikes physical contact' and 'does not look up when spoken to'). These results suggest a dimension related to reduced social interaction that is not specifically related to the Internalizing spectrum, but also not fully captured by the 'Broader Autism Phenotype'. This is very much in line with the idea that the social and non-social aspects of BAP should be studied as separate traits rather than as a single dimension (Happe & Ronald, 2008). Reduced social interaction may be specifically relevant to understand comorbidity with Social Anxiety Disorder and may also be related to possible (future) comorbidity with depressive disorders.

Second, the scale Fear of Change loaded on the INT-factor. This subscale consists of only three items, which have low prevalence in the general population and are related to strong emotional reactions to change, which is typical for some children with ASDs. A speculative hypothesis derived from this finding is that these strong reactions to (social) change underlie or are risk factors for Internalizing problems.

BAP and the Externalizing Domain

Children with ASD often meet criteria for Oppositional Defiant Disorder (ODD), and less often for Conduct Disorder (CD; de Bruin, et al., 2007; Simonoff, et al., 2008). The other way around, ASD-problems are highly prevalent in children with ODD or CD (Gilmour, Hill, Place, & Skuse, 2004; Luteijn, et al., 2000). The current study suggests one specific source that may contribute to this comorbidity.

The scale Not Tuned loaded on EXT. This can be explained by the fact that the items of this scale are clearly related to aggressive and disobedient behaviors. Nevertheless, the scale contains items on a specific kind of problems that are not fully captured by EXT and the CBCL-scales, as the scale also loads on the BAP-factor. The hypothesis that can be derived from this is that the subscale reflects a source of disobedience and opposition that is specifically related to social deficits, difficulties in

social information processing or strong reactions to change. The question to be answered is whether overlap between problems from the syndromes ODD/CD and ASDs is specifically related to the factor Not Tuned and what underlying mechanisms influence this specific covariance. It may be relevant for understanding and diagnosing Externalizing behaviors to distinguish between children who have a lot of social deficits and poorly understand social conventions and those who don't, because they may differ with regard to their motives and triggers for aggression and opposition, and in their response to interventions (Gilmour, et al., 2004).

BAP and ADHD

Children with an ASD often meet criteria for ADHD (de Bruin, et al., 2007; Simonoff, et al., 2008). The other way around, autism symptoms are often found in children with ADHD (Bishop & Baird, 2001; Mulligan, et al., 2009). To understand this phenomenon it may be important to distinguish between two aspects of ADHD: attention problems on the one hand and impulsivity/hyperactivity on the other (Ronald, et al., 2008). This distinction was found important in a recent study by Lahey, et al. (2008) and has been supported by a recent Latent Class Analysis by Volk, Henderson, Neuman, and Todd (2006), which distinguished three types: Severe Inattention, Severe Combined and Mild Combined. The results of the current study suggest two specific sources of co-occurrence between ASD-symptoms and ADHD.

First, the scales Attention Problems and Orientation both loaded strongly on the same factor of the bi-factor model (F4* in figure 4). Similarly, in an EFA by Hartman et al. (2006) items from these scales loaded on the same factor. This points to the possibility that one specific source of comorbidity between ADHD and ASD-symptoms is related to attention, executive functioning, and information processing difficulties. Children with ADHD and PDD-NOS have been shown to perform very similarly on a number of tasks measuring executive functioning (Gomarus, Wijers, Minderaa, & Althaus, 2009). The factor F4* may be specifically related to these problems in executive functioning and to the 'inattention' aspect of ADHD, which may constitute as specific subdomain of problems that can be distinguished from BAP, INT and EXT.

Second, the CBCL-scale Attention problems does not capture the 'hyperactivity and impulsivity' component of ADHD. Items regarding this component are found in the Aggressive Behavior scale. This scale loads on the EXT factor and so does the Not Tuned scale of the CSBQ. This suggests that a second source of co-occurrence between ADHD and ASD-symptoms may involve a specific Externalizing feature of both syndromes. This idea is very similar to the hypothesis we developed for the overlap between ODD and ASD-symptoms and results in a more general hypothesis that similar underlying (EXT-related) mechanisms may be implicated in the comorbidity between these three syndromes. One such common mechanism may involve deficient social problem solving (Matthys, Cuperus, & Van Engeland, 1999), which can increase the chance to get into conflicts with others.

Implications for diagnosis and clinical practice

Understanding the latent structure underlying multiple problem domains can be crucial for advancing diagnostic systems over time. The current study points to some important ways in which clinical diagnosis may be improved.

First, measuring BAP and its subdomains may not only be relevant in the context of diagnosing ASDs, but also in the context of other, more prevalent, DSM-IV diagnoses. The current study shows that treating BAP as a completely distinct domain of psychopathology may neglect important relations between BAP-subscales and other domains of psychopathology. It is counterintuitive to assume that ASD-symptoms are only relevant if severe enough to meet criteria for a DSM-IV syndrome.

Second, the study provides a method to eliminate the unsatisfying heterogeneous residual category PDD-NOS. From the bi-factor model a stepwise approach to diagnosis may be derived: First, a diagnostician would use a measure of non-specific amount of problems. Second, he or she would investigate scores on specific problem domains (for example, INT, EXT, Attention, Reduced Contact, BAP). Third, in some cases, a more detailed analysis of the subscales of these domains may add specific diagnostic information. In such a diagnostic procedure that follows a dimensional model of common and specific features, the category PDD-NOS can be replaced by a broad BAP-dimension with more specific subscales. This approach, which is similar to a recent proposal of Watson (2005) regarding the Mood and Anxiety Disorders, has several advantages, among which (a) decreased artificial comorbidity, (b) more attention to both general and specific aspects of diagnosis, (c) an emphasis on the gray areas between current categorical syndromes and thereby (d) a closer link to the covariance structure of problems in the general population. Furthermore, if categories are preferred for clinical practice they can be derived on the basis of carefully chosen cut-points on a dimensional structure (e.g. Kamphuis & Noordhof, see chapter 3).

Third, the results are in line with a recent tendency in research and clinical practice to place more emphasis on trans-diagnostic processes as opposed to focusing on predefined syndromes. Even if syndrome boundaries would "neatly carve nature at its joints" (Waller, 2006), underlying processes and successful interventions may be shared between multiple problem domains. Since it is well-know that the actual situation is one of overlapping domains and gray arrays between them the issue of trans-diagnostic processes is even more relevant and neglecting it can result in suboptimal treatment.

Conclusion

This study shows that (a) problems traditionally related to the domain of ASDs can be adequately integrated into general population based dimensional models of psychopathology, (b) the 'Broad Autism Phenotype' can be regarded as a specific domain of problems that can be distinguished from the domains of Internalizing, Externalizing and Attention Problems, and (c) specific subdomains of BAP are differently related to INT, EXT and Attention Problems.

Chapter 3

On Categorical Diagnoses in DSM-V: Cutting Dimensions at Useful Points? *

Jan Henk Kamphuis, Arjen Noordhof

Abstract

DSM-V will likely place more emphasis on dimensional representation of mental disorders. However, it is often argued that categorical diagnoses are preferable for professional communication, clinical decision-making, or distinguishing between individuals with- and without a mental disorder. For these specific aims, utility-based categories can be created on the basis of a dimensional framework by using cut-points. This article addresses several ideas for combining categorical and dimensional approaches like prototype matching, adding scores of symptom-severity, and introducing utility-based categories in dimensional models. We identify alternative objectives for specifying cut-points and describe ways of determining the cut-points accordingly. It is recommended that for creating standard diagnostic concepts fixed cut-offs be used, as this promotes accumulative science, but these cut-offs may not be optimal for other clinical decisions, because of local base rates and decision-specific (dis-)utilities. ROC-curves can facilitate the comparative evaluation of the trade-off between sensitivity and specificity for multiple cut-points and diagnostic rules. We advocate a DSM-V that contains both categories and dimensions in order to serve the multiple and complex aims of utility and validity.

Introduction

DSM-V is likely to place more emphasis on dimensionality than the current DSM-IV system. There are good scientific reasons to go this way. First, evidence regarding the nature of psychopathology supports dimensional models for various groups of disorders (e.g. Haslam, 2003). Second, the rampant but decidedly not random comorbidity associated with the fourth edition of the Diagnostic and Statistical Manual for Mental Disorders (4th ed.; American Psychiatric Association, 1994) suggests that more fundamental dimensions underlie the current classes of disorders. Specifically, based on structural equation modeling of covariance between common psychological disorders, Krueger (1998) proposed two spectra of psychopathology (i.e., internalizing and externalizing psychopathology) with distinct underlying etiology (Krueger & Markon, 2006a). Similar dimensions were derived from a bottom-up analysis of

* *This chapter was published as an article in 2009 in Psychological Assessment, Vol 21(3), pg. 294-301.*

symptoms of child psychopathology (Achenbach, et al., 2008). Third, dimensional systems have the psychometric advantage that more statistical power is retained for detecting discriminations in subsequent analyses (as argued for example by Frances, 1993; or Widiger, 1992).

Proponents of categorical systems on the other hand assert that (clinical) practice favors categories. Putative traditional advantages of categorical systems include a) greater ease of communication (i.e., clinicians prefer category names to profiles of scores on dimensions), b) continuity with current clinical practice and clinical decision making (e.g., admission or not; antidepressant medication or not), c) greater ease for counting purposes, and d) better fit with reimbursement policies. Verheul (2005) recently argued that most of these traditional advantages are in fact minimal, and that smart dimensional models can (learn to) accommodate the various purposes of diagnostic systems just as well or better. Moreover, Clark & Harrison (2001) argued that these perceived advantages are essentially un- (or anti-) scientific in nature.

That said, categories may serve the practical aims of specifying the scope of clinically significant psychopathology in need of treatment, and of creating boundaries between syndromes in order to facilitate communication. As will be argued in the current paper, utility-based categories can be created within a dimensional classification system. From this perspective, the question is not so much 'which cut-offs provide valid distinctions between disorders?', but rather 'how to create and evaluate cut-offs that best serve the complex aims of a diagnostic system?' To answer this question, we will distinguish between three aims for which the DSM-system is used: a) developing standard international concepts of psychopathology, b) making the expert-decision of whether a set of symptoms should be regarded a mental disorder or not, and c) making predictions and decisions about treatment and risky outcomes. With these three issues in mind, we will discuss approaches for developing cut-off points for categorical diagnosis. Specifically, should we use flexible cut-offs, adjustable for local base rates and/ or decision-specific profiles of disutility? How to decide on a specific optimal score for cut-offs? Is it useful to use multiple cut-offs rather than just one? This brief review will discuss these questions and in so doing touch on the potential contribution of Signal Detection Theory techniques for selecting cut-points, the pros and cons of prototype matching, and a recent proposal that advocates supplementing the traditional syndrome-based DSM-system with ratings of symptom-severity (Helzer, Kraemer, & Krueger, 2006). We will limit our discussion strictly to the issue of boundary setting and will treat the criteria sets as givens; for a thoughtful discussion of how diagnostic criteria might be specified as harmful dysfunction indices in order to minimize diagnostic error, the reader is referred to Wakefield and First (2003). To illustrate the abstract issues that will be discussed, we will use examples from the internalizing spectrum, but the arguments apply to other domains of psychopathology as well.

Latent Structures of Psychopathology

The primary aim of psychiatric diagnosis is to provide information about the conditions (i.e. latent causal structure) from which psychological problems emerge. A formal diagnostic system like DSM-IV provides concepts (e.g. Major Depressive Disorder) that can be diagnosed by applying a specific set of diagnostic rules. The validity of such a system ultimately depends on the question whether variance in diagnoses is in fact caused by variance in kinds of psychopathology to which the diagnostic concepts refer (Borsboom, et al., 2004). Of course, such a causal relation between diagnoses and actual psychopathology cannot be observed, so diagnoses can be regarded as hypothetical constructs (Cronbach & Meehl, 1955). The validity of diagnostic systems can be evaluated on the basis of a process of construct validation, which amounts to simultaneously testing measures of psychological constructs and the theories of which the constructs are a part (Strauss & Smith, 2009).

One of the methods that can be employed for construct validation is latent variable modeling on data from general population samples. An important finding from these studies is that many syndromes of DSM-IV are suboptimal because the hypothesis that they represent natural categories is probably false (Haslam, 2003; Kubarych, Aggen, Hettema, Kendler, & Neale, 2005; Lubke & Neale, 2006). Natural categories (or taxons; Waller & Meehl, 1998) are non-arbitrary types (e.g. gender, species, or lung-cancer). In psychopathology natural categories are not easily demarcated just on the basis of observation or intuition and sophisticated statistical methods have been proposed to discover latent natural categories underlying the observations (Meehl, 1992). These techniques are, among others, taxometrics (Ruscio, 2009) and Latent Class Analysis (Lubke & Neale, 2006).

To illustrate this issue of suboptimal categories we will briefly discuss the DSM-IV categories of 'Major Depressive Disorder' (MDD) and 'Generalized Anxiety Disorder' (GAD). These syndromes are conceptualized as two distinct categories that belong to the domains of mood disorders and anxiety disorders, respectively. In general, research has not supported a sharp boundary between the two conditions, while both behavior-genetic and phenotypic analyses indicate that symptoms from these syndromes have much more shared than specific variance (Mineka, Watson, & Clark, 1998). Furthermore, taxometric and latent class analyses do not support categorical models for either 'Major Depressive Disorder' or 'General Anxiety Disorder' (Haslam, 2003; Kubarych, et al., 2005; Lubke & Neale, 2006). Given that current evidence does not support the idea of two distinct diseases at all, dichotomous diagnostic rules will very likely result in loss of information and artificial comorbidity due to arbitrary boundaries (Kraemer, Noda, & O'Hara, 2004).

An appealing alternative to categorical syndromes is provided by models in which psychological problems are hypothesized to exist along continuous dimensions. Such structures have been developed and replicated with the use of factor analysis (Krueger, 1999). In particular hierarchical factor analysis has proven useful in distinguishing factors that are common to multiple syndromes and factors related to a

specific subset of problems. For example, Watson (2005) argued for a structure of the emotional disorders (cf. internalizing spectrum) that is much closer to current empirical knowledge than the DSM-IV conceptualization. In this structure GAD and MDD are both grouped into a higher order domain labeled 'Distress disorders', which is distinguished from 'Bipolar disorders', and 'Fear Disorders'. In this model, the large shared variance of GAD and MDD is captured as indicative of the common factor 'Distress disorder' rather than manifested as comorbidity between two distinct disorders. (Watson, 2005)

It has been argued that categories should not be regarded unscientific and that open-mindedness to the scientific advantages of both categorical and dimensional representations is preferable to disregarding either of them (Pickles & Angold, 2003). Nevertheless, in absence of evidence for natural categories, dimensional factors will often be more efficient in terms of power, retaining information and reliability in comparison with the somewhat arbitrary dichotomizations of DSM-IV. Moreover, constructing categories on the basis of dimensional information is easier than the other way around.

Latent Factor models can be directly compared with Latent Class models by comparing indices of model-fit (Lubke & Neale, 2006). Furthermore, the technique of factor mixture modeling offers the possibility to test models that contain both dimensions and categories (Lubke & Muthen, 2005). These techniques can provide a basis for introducing categories after adopting a general dimensional diagnostic framework, provided that there is enough support for the existence of categories and enough indicators to reliably predict to whom they apply.

Creating Utility-based Categories in a Dimensional Framework

In our view the above described approaches currently provide the most rational basis for developing an empirically informed classification of psychopathology. However, diagnosis is not exclusively a matter of maximizing the validity of classification. There may be good, though maybe not strictly scientific, reasons to create non-natural diagnostic categories on the basis of a dimensional classification. The purpose of such categories will not be 'to cut nature at its joints', but to provide useful tools for those who use the diagnostic system. In the following we will discuss two alternative approaches to attain this goal.

Using Prototype Matching

Diagnosis can be based on the use of Prototype Matching (PM). In PM, diagnosis does not flow from explicit diagnostic (counting) rules. Individual specific diagnostic criteria are replaced by a global description of a prototypical patient who fits a particular diagnosis (e.g. Shedler & Westen, 2004). Specific criteria can be incorporated in these descriptions, but these are not rated as individual items. As such, the prototype can be considered a single, complex item that is scored on a 5-point rating-

scale (instead of a more simple and dichotomously scored criterion, as is current practice in DSM-IV). Clinicians determine the overall resemblance or match between patients and prototypic descriptions. This procedure does not constitute a full return to the DSM-II vignettes, as the comprising elements can now be empirically derived, and rated on a five-point rather than dichotomous scale (Spitzer, First, Shedler, Westen, & Skodol, 2008). PM shows good utility in that practitioners seem to prefer PM over other diagnostic approaches, including the DSM rendering of information (Spitzer, et al., 2008), perhaps because of better fit with their naturally occurring decision making process. This advantage should not be taken lightly, as user-friendliness and acceptability are important determinants of the extent to which a diagnostic system will be (faithfully) adopted in clinical practice[4]. Moreover, PM also has some apparent important drawbacks. First, the validity of the specific combination of symptoms that constitute each prototype cannot be evaluated when the symptoms themselves are not rated. Second, accumulation of scientific knowledge may be rendered more difficult with a PM approach to diagnosis, as the covariance structure of symptoms cannot be evaluated. Data on the covariance structure of alternative criterion sets is now a primary avenue for furthering our understanding of the structure of psychopathology (e.g. Krueger, 1999; Widiger & Clark, 2000). PM thus introduces a black box in that it remains implicit what aspects of the global description are guiding the clinicians' ratings. This may easily result in a conservation of expert bias, especially when clinicians are no longer required to specify how they derive their diagnoses. Clinicians may differ strongly in their weighing of the comprising information units and such individual differences are likely to be detrimental to the reliability of diagnosis. In sum, prototypes may serve to construct a common international standard for diagnosis that is preferred by many practitioners, but PM does not seem optimally geared for the scientific objectives of the DSM.

Enhancing DSM with Symptom-severity Ratings

A second, in our mind more promising approach, is to scale individual criteria to some extent and to then use a sumscore to describe symptom severity. A recent, quite practical proposal in this vein, originates from the domain of substance use disorders (Helzer, Kraemer, et al., 2006; Helzer, van den Brink, & Guth, 2006). These authors proposed supplementing categorical substance use criteria with a dimensional quantitative component by having patients rank each criterion on a simple three-point scale running from 0 = not present, 1 = mild, to 2 = severe, thus both providing a patient with a convenient intermediate between 'yes' and 'no', while also yielding the desired quantification at the symptom level. Hence, their diagnostic program entails including *both* categorical and dimensional representations of diagnosis by introducing

[4] Enhanced clinical utility is by no means unique to the prototype matching approach. A study by Samuel and Widiger, for example, showed that clinicians judged Five-Factor Model based descriptions of cases of greater clinical utility than diagnostic categories (Samuel & Widiger, 2006).

severity ratings for individual (DSM-V) symptoms. Such combined scoring will allow for empirical comparisons of the heuristic utility of both systems, and experimentation, with additional symptoms not reflected in the categorical system. The authors state it is vital that the dimensional and categorical systems be communicative; i.e., that the quantitative scale can be translated into categories. The final step would involve relating the dimensional scores to the categorical diagnostic threshold by some algorithm, to yield a score that is identified with meeting traditional diagnostic criteria. DSM-V wide adoption would help solve the co-morbidity issue by substituting profiles of symptom severity across disorders for the current multiple diagnoses.

To illustrate, consider the Major Depression Episode diagnosis. To meet diagnostic criteria, during at least a two-week period the patient should report significant depressed mood and/ or anhedonia, and an additional three (out of seven remaining) criteria, (as well as satisfying the B, C, and D criteria). In addition to making nominal decisions for each symptom, the clinician would rate each of these symptoms on the 0 = not present, 1 = mild, to 2 = severe scale suggested by Helzer, Kraemer, & Krueger (2006). The dimensional rating would thus have a range of 0 to 27, and logistic regression, recursive modeling techniques, or other statistical tools may have established that a score of 14 (or 15, or 16, etc.) on the dimensional ratings as optimally predictive of the categorical major depression diagnosis for a range of pertinent base rates. This practice would not require significantly more effort from the clinician than the current DSM-IV more subjective appraisal of severity.

In the proposal of Helzer, Kraemer, & Krueger (2006), the syndrome structure of DSM-IV is basically retained. However, DSM-V wide adoption would also offer the possibility of analyzing the symptoms from the different categories together. As described above, latent variable modeling has been used to develop dimension that approach the structure of psychopathology in general population samples better. Subsequently, utility-based categories can be created by determining cut-points on this newly found dimensional structure.

Aims and Types of Utility-Based Cut-points

Utility-based categories can be created by specifying some diagnostic rule that combines multiple indicators to select a cut-point to make a dichotomous decision about the presence or absence of disorder. Such rules may describe a simple cut-point on a scale score, for example a score of 14 (or 15, or 16) on a severity-scale for Major Depressive Disorder, but may also involve a specification of essential and/ or polythetic criteria that should be present (e.g. at least two core-symptoms of depression). In the following, we distinguish between scenarios in which a) a cut-point is defined by a fixed norm of statistical deviance, b) multiple cut-points are defined to indicate different levels of symptom-severity, c) a cut-point is defined to facilitate the expert-task of deciding whether a set of symptoms is to be regarded as a disorder (the traditional 'diagnostic' cut-point), and d) a cut-point is defined to provide a basis for statistically informed decisions regarding treatment or prevention.

Cut-points Based on Statistical Deviance

A quick-and-easy approach is to simply choose a cut-point at a certain percentile of scores on a scale in a representative sample of the population of interest. Individuals who score above that percentile are diagnosed, individuals below the percentile are not. The drawback of this approach is that it fixes the expected amount of patients in a sample without clear criteria to support or reject this choice. However, if the aim is simply to diagnose a certain percentage of the population this may not be a problem (e.g. 2.3% of the population is regarded 'highly gifted' on the basis of IQ-scores; a constructivist concept).

Severity Cut-points

It is not necessary to choose between dimensional scores and dichotomous classes. Relying on scale-scores for communicating information to patients and colleagues may prove difficult, but natural language does offer multiple alternatives for capturing a latent dimensional structure. For example, distinguishing between mild, moderate and severe depression may make more intuitive sense then either communicating a score or the presence/ absence of a diagnosis. By setting a number of severity cut-points in DSM-V, it would be possible to create a useful standardized language that is closer to the dimensional latent structure.

A 'Diagnostic' Cut-point?

The DSM-system does not function as a mirror brightly reflecting the nature of insanity. Instead, DSM-IV demarcates important cultural boundaries with large consequences for the individuals involved. Cut-points and specific criteria are tools by which clinical experts judge whether a set of problems is to be regarded a disorder. For some diseases, like lung cancer, there is a well-understood causal relation between symptoms and disease and a 'gold-standard' means of ascertaining whether the disease is indeed present. In those cases the quality of expert-judgment is a purely technical issue, because there is a criterion to measure whether the expert-judgment reflects the true state of nature.

The situation is radically different for utility-based categories that are developed in the absence of natural categories and 'gold-standard' measures. In this situation, which is the rule in psychiatry, the information on which a diagnosis is based can be improved by developing better measurement instruments and by the process of construct validation, but if the underlying problems are of a dimensional nature this will not result in an empirical cut-point. Therefore, the distinction between normative daily problems and mental disorders is ultimately based on human judgment. In current Western societies clinical experts are expected to make the distinction between normative problems and problems that are regarded disorders. Often these decisions are made in suboptimal conditions on the basis of a limited amount of imperfect information. Furthermore, research on human expertise provides ample support for cognitive limitations and biases involved in expert-judgment (Dawes, Faust, & Meehl,

1989). Finally, in the absence of an international standard it can be expected that expert-judgments will vary over different clinical settings and geographic locations and that this divergence will increase over time.

Hence, on balance, we believe that it may be wise to create an international standard 'diagnostic' cut-point to facilitate expert-decisions and argue for a decision-making process that is rooted in empirical classification and that is sensitive to the consequences of creating cultural boundaries. This sensitivity to consequences is a difficult ethical issue for two reasons. First, the consequences are related to a wide variety of interests and individuals, including patients, clinicians, students and researchers, mental health services, health insurance companies, government reimbursement and prevention programs, etc. Second, the consequences are to a large extent unpredictable and uncontrollable after the diagnostic cut-point has been set. While those who make the distinction may be well aware of its inherent limitations, it often proves difficult or even impossible to fully communicate this understanding to a wider audience.

A clinical cut-point should be based on all information that is relevant to the decision whether a set of problems should be regarded a clinical syndrome. The DSM-IV concept of mental disorder includes the requirement that the 'symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.' This statement implies a categorical decision: the patient does or does not meet this requirement, and the assessment is crucial for whether or not a set of problems is considered a disorder. This decision depends on a number of considerations that are, to some extent, external to the symptoms of a narrowly defined mental disorder itself: daily life functioning and quality of life, 'need for treatment', or danger to self or others. As argued by Kessler (2002), external analyses based on epidemiological studies may aid in providing an empirical basis for evaluating the multiple criteria on which a clinical cut-point can be made. External analyses seek disjunctions in the gradients for external correlates of psychiatric morbidity, which include variables external to the diagnosis itself, such as comorbidity, family history, impairment (for a discussion of psychiatric validators, see Robins & Guze, 1970). Optimal diagnostic thresholds may not prove to be consistent across external correlates, but a pattern may predominate. In the end, experts, informed by the results of the external analyses, field trials, and their personal experience and expertise, make the cut.

Cut-points in the Context of Prediction and Decision-making

DSM diagnosis may also be used to support other important clinical, dichotomous decisions: e.g., to admit a patient to an inpatient unit, to prescribe antidepressant medication, or other decisions that broadly concern treatment selection, or risk prevention. There is no principal reason to assume that those concepts that are most useful for developing an international standard for communication about psychopathology that meet certain criteria of sensitivity and specificity, are also the

best concepts for making a wide variety of clinical predictions. Moreover, the optimal criteria and cutoffs for making decisions are likely to differ between decisions. For example, screening will require more lenient cut-points than sampling prototypical cases (Robins & Guze, 1970). Hence, the DSM syndromes (and the cut-points they are based upon), are exceedingly unlikely to be consistently optimal for the great diversity of everyday clinical decisions encountered by clinicians. Nevertheless, efforts toward standardization may be helpful as a convincing body of evidence suggests that actuarial judgment often outperforms clinical judgment (Dawes, et al., 1989).

Setting and Evaluating 'Diagnostic' Cut-points

To evaluate the performance of a diagnostic system one can empirically relate the outcomes of diagnostic rules to a criterion of what the diagnosis should be (Swets, 1988). For the purpose of creating categories based on human judgment, a 'LEAD'-standard may be developed (Spitzer, 1983). LEAD (Longitudinal, Expert, All Data) involves creating optimal circumstances (e.g., relying on clinicians who have demonstrated their reliability, and including information from multiple sources, and from multiple time points) for expert judgment and using these judgments as a criterion to evaluate diagnostic rules. What circumstances are optimal is a complex question beyond the scope of the current paper. However, in our view it is crucial: (1) that judgment be based on all information relevant to the expert decision and no irrelevant information, (2) that the information be derived from well-validated instruments whenever possible, and (3) that multiple experts judge independently in order to empirically investigate the amount of consensus.

Subsequently, this LEAD-standard can be used to statistically evaluate the performance of multiple diagnostic cut-points. To this end a sample of subjects is needed for whom all criteria are measured on which the DSM diagnostic rules will be based, and for whom the LEAD-standard judgment is known. The results of the diagnostic rule and the criterion can be summarized in a 2 x 2 Contingency Table (see Table I). By definition, inaccurate diagnostic decisions fall in two categories (see the Contingency Table, Table I): one may diagnose a person who should not be diagnosed according to the LEAD-standard (Cell B, False Positive) or one may not diagnose a person who should be according to this standard (Cell C, False Negative). Their complements, the so-called hit rates, can be divided into row- and column indices. Column indices include the proportion of accurate positive diagnoses (Sensitivity, or a/[a + c]), and the proportion of accurate negative diagnoses (Specificity, or d / [b + d]). Again, it should be emphasized that in this case the conventional term accurate does not refer to truth, but to approximation of the LEAD-standard.

Table 1. *Contingency Table, Crossing Criterion (e.g., LEAD Expert Opinion, in Columns) and Predictions from Diagnostic Rule (Rows).*

		Criterion: e.g. LEAD Expert opinion	
		'Present'	'Absent'
Diagnostic Rule	*Test score > cut off:* 'Diagnosis'	Cell A True positive (a)	Cell B False positive (b)
	Test score <= cut-off: 'No Diagnosis'	Cell C False negative (c)	Cell D True negative (d)

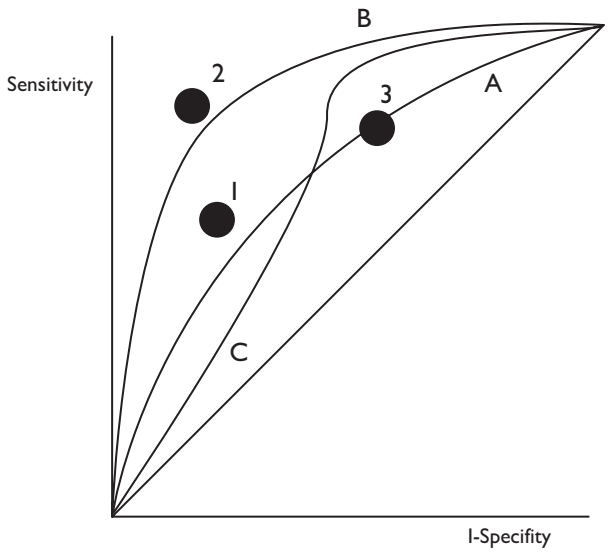

Figure 1. Receiver Operating Characteristics (ROC)-plane with some examples of ROC-curves and points.

A particularly useful tool that can be used to evaluate multiple contingency tables that result from multiple diagnostic rules is rooted in Signal Detection Theory (Swets, 1988)[5]. Specifically, the Receiver Operator Characteristics curve, better known as ROC-curve (McFall & Treat, 1999; Murphy, et al., 1987) creates a two dimensional plane that plots the trade-off between sensitivity and specificity. As illustrated in Figure 1, each diagnostic rule (i.e., cut-point) results in one point (see 1, 2, and 3) in the ROC-plane. If a number of different cut-points are considered for the same scale, this can be plotted as an ROC-curve (see A, B, and C) that connects the results for the different cut-points. The diagonal in figure 1 represents diagnosing on the basis of tossing a fair coin. All points above that line represent some increase in diagnostic accuracy.

When comparing points in an ROC-plane, there are two possibilities. Diagnostic rules may outperform each other (e.g. Figure 1: point 2 versus point 1, or curve B versus curve A) in which case one point or curve is clearly superior. Alternatively, a diagnostic rule may result in more 'True Positives' (i.e. diagnostic rules and LEAD-standard both result in a positive diagnosis), but also more 'False Positives' (i.e. diagnostic rule results in a positive diagnosis, but LEAD-standard does not). In that case the utility of true positives should be weighed against the disutility of false positives, which follow from the aim of the diagnosis. When comparing multiple cut-points on the same dimensional score there is always such a trade-off, because a lower cut-point will inevitably result in increased numbers of both 'True Positives' and 'False Positives'. When comparing different sets of criteria and/ or different measurement instruments this is not necessarily the case.

Different diagnostic rules will result in different combinations of rates of sensitivity and specificity and it will be necessary to choose a specific trade-off between types of error; which inevitably requires weighing the (dis-)utilities of diagnosis and non-diagnosis. It is of specific concern if diagnostic rules are so strict that subjects in need of treatment according to current expert consensus are not diagnosed by the diagnostic rule (i.e. 'False Negatives'). However, too permissive diagnostic rules may result in pathologizing normative daily-life trouble. Of note, DSM itself advocates some degree of flexibility in the use of its diagnostic rules, and explicitly recognizes the value of clinical judgment in special cases.

The advantage of evaluating diagnostic rules on the basis of Sensitivity and Specificity or ROC-curves is that these are independent of the base-rate of disorders in a particular sample. However, in clinical practice these measures may not answer the most relevant question with regard to diagnosing individuals. When giving (or not giving) a diagnosis on the basis of a diagnostic rule, a clinician is probably more interested in knowing the probability that this diagnosis fits the conventional diagnosis.

[5] *Signal Detection Theory is certainly not the only statistical technique available. Notably, Bayesian statistics offers multiple tools to evaluate the performance of diagnostic systems and to formally weigh the multiple utilities and disutilities of over- and underidentification. However, these technical issues are beyond the scope of this article.*

This chance is represented by the row indices of the contingency-table: the proportion of True Positives (Positive Predictive Power, PPP, or a/ [a + b]), and the proportion of True Negatives (Negative Predictive Power, NPP, or d/ [c + d]), with 'True' as before, defined by utility-based cut-points. While the column indices are independent of base rate, it can be readily seen from the contingency table that the PPP and NPP indices are highly dependent on the base rate (i.e., a + c/ N) (Baldessarini, Finklestein, & Arana, 1983; Kamphuis, Finn, & Butcher, 2002; Meehl & Rosen, 1955; Streiner, 2003). To evaluate a diagnostic system, PPP's and NPP's may be estimated for multiple cut-points and multiple base-rates. From such analyses one might select one single cut-point that has a satisfactory PPP and NPP for a range of different base rates reflecting standard clinical practice.

Alternatively, one might use flexible cut-points that maximize the diagnostic accuracy for each setting (i.e. based on local base rate), much like we adjust levels of 'clinical significance' for a psychological test in different settings (e.g. higher cut-points for social desirability in custody evaluations; higher cut-points for somatic complaints in geriatric settings) (Finn, 1982). Different patients exhibiting the same symptoms may then receive different diagnoses, as the local circumstances dictate different optimal cutting scores. Regarding the influence of base rates in setting diagnostic thresholds for symptom counts, Grove (1985) demonstrated, using Monte Carlo simulations with parameter values similar to those for the DSM decision process, that bootstrapping diagnoses for local base rates does not improve diagnostic accuracy in clinically meaningful ways. On the other hand, when base rates become more extreme and group separations smaller, adjusting cutting scores may improve diagnostic accuracy (Hsu, 1988). A second argument that has been put forward in favor of using flexible rather than fixed cut-offs is that Type-I and Type-II errors (errors of overidentification versus underidentification) may weigh differentially in different circumstances, depending on the consequences that such errors have. Some disorders, for example, might have very high labeling costs, which would argue for higher disutility of False Positive errors. For similar reasons cut-offs for diagnostic decisions might even be individualized, according to Finn (1982, 1983). Widiger (1983) forcefully argued for invariance of diagnostic cut-offs for considerations external to the diagnosis itself (such as idiosyncratic [dis-]utilities of receiving the diagnosis). The main thrust of his argument, i.e. that accumulative science is only possible when the same label refers to patients exhibiting similar sets of symptoms, seems well-taken. If the aim is to use categorical syndromes as a basis for cumulative science, each categorical judgment should be based on the same criteria and the same cut-offs. For clinical decision making beyond standard diagnostic classification however, Finn's point makes good sense. For example, in selecting an optimal dose of psychosocial or psychomedical treatments for a particular patient, clinicians may use different severity criteria than the standard cut-off implied in the formal diagnosis.

Conclusion

The DSM-IV diagnostic system has become a widely-used guide for clinical practice, as well as for decision makers from mental health policy and insurance domains. The aims of this diagnostic tool are more complex than merely providing maximally valid representation of diagnostic information. Three principal other aims include a) ease of communication, b) deciding which individuals will be regarded as suffering from a mental *disorder*, and c) facilitating clinical decision making in general.

Our analytic review can be summarized as follows. We believe that structural modeling approaches on general population data have significantly advanced the knowledge of the latent structure of psychopathology. Specifically, hierarchical factor models provide a useful tool to distinguish between common and specific features of psychopathology and restructure the diagnostic system on the basis of empirical knowledge. Furthermore, we have described techniques that allow for detection and introduction of natural categories within a dimensional framework, and we hold that only after a dimensional framework has been established, utility-based categories should be introduced. In general, it is our view that dimensions likely provide more valid descriptions of psychopathology and that cutting scores should be imposed on dimensions where useful. In closing, we will present a specific proposal as to how DSM-V might handle such cut-points.

We believe the previously discussed Helzer, Kraemer, & Krueger (2006) proposal provides a realistic starting point for restructuring the DSM, that combines several important features. It entails supplementing the DSM-V diagnostic criteria with a three-point severity scale, preferably across diagnoses. As such, it provides continuity with the current system, and in our judgment, is sufficiently user-friendly and intuitive to expect widespread adoption. Compared to the current DSM-IV-TR severity ratings for disorders, the Helzer et al. (2006) proposal offers two advantages: it includes *symptom* severity ratings, instead of the more or less global *syndromal ratings*, and b) it derives the overall symptomatic severity index by an objective bottom-up algorithm (i.e. a sumscore over the comprising symptom severity ratings). The simple three-level format will likely promote clinical adoption and thus facilitate systematic clinical judgment of these ratings. Moreover, by introducing dimensional ratings, it opens the door to more powerful statistical analyses that can further elucidate the latent structure of psychopathology. In that way, the proposal is more open and dynamic than the current system, and more likely to over time approximate the true structure of psychopathology.

In addition, we would like to introduce two (or more) severity cut-points. Based on the sumscore of the severity ratings, it seems instructive to suggest two additional cut-points that mark a 'moderate' and a 'high' level of symptomatology. This practice would serve as a reminder for users of the system that they are not dealing with natural categories, and that different cut-points may be appropriate for different aims. These severity-descriptors can be given for multiple diagnoses without suggesting the presence of multiple comorbid diseases at the same time. A more daring revision,

supported by the latent modeling approaches, would be to use severity-labels at different levels of specificity. For example, a patient might be described as suffering from a 'moderate' level of Emotional Distress, and specifically from a 'severe' level of GAD-symptoms. In our view, these are intuitively plausible natural language descriptors that are well-suited for communication with colleagues and patients and more apt than the current practice of either underdiagnosing comorbidity or suggesting that many patients suffer from a series of distinct mental disorders. However, it remains an open-ended empirical question whether clinicians can effectively use such a system.

For what is typically referred to as the 'diagnostic' cut-point, a single, fixed cut-off would serve the needs of various decision makers best. This cut-point is likely to fall between the 'moderate' and 'severe' cut-points previously discussed. To compare multiple diagnostic rules, we suggest the use of statistical techniques (e.g., ROC-curves) and a criterion of optimized, LEAD-standard expert judgment. Simple diagnostic rules may be created on the basis of cut-points on either symptom counts or severity scores. The latter may have the advantage of being more statistically discriminating, while the former is more akin to current practice. As it is more fine-grained, opting for the sumscore of symptom severity ratings would probably also yield less dramatic (and embarrassing) shifts in prevalence when the diagnostic threshold is slightly adjusted over time (Regier, et al., 1998). However, this would be a more radical step, and would foreclose on the opportunity to directly compare the value of the current dichotomous format with the symptom severity-ratings.

For several clinical decisions that may involve prediction, the judicious use of flexible cut-points (Kamphuis, et al., 2002) that takes into account local base rate information and the type of decision at hand may be indicated. Using ROC curves, it is easy to demonstrate that different cut-offs are associated with different profiles of sensitivity and specificity, and that decision makers can choose a cut-point that best matches the profile of respective disutility for False Negative versus False Positive errors. In contrast to the purposes implicated for the 'diagnostic' cut-point, in these scenarios there often is less need for uniformity or consistency across settings.

In the current article we have made a sharp distinction between the DSM-V aims of representing information on the one hand and making diagnostic decisions on the other. However, we certainly do not argue for the development of two separate diagnostic systems. On the contrary, in our view science and practice would profit from a communicative system with both elements. Scientific progression with regard to the latent structure of psychopathology should be adopted into our diagnostic systems, which therefore need to offer the flexibility of restructuring classification on the basis of robust empirical evidence. At the same time diagnostic rules in clinical practice and the common language of experts and patients profits from a certain amount of conservatism. Stability in diagnostic language is crucial for the development of communication among experts and between experts, patients and the general public. For this communication to be meaningful it needs to be rooted in a solid,

empirically supported system of classification, which with the current state of science will contain a large amount of dimensionality.

To conclude we would like to endorse the sentiment expressed by Frances et al. (1991), who stated that 'the highest purpose of the DSM-IV is that it encourage and facilitate the research that will render it obsolete' (Frances, et al., 1991; p411). In our view, much the same should be true for DSM-V, and supplementing (multiple) utility-based categories with dimensional information may go a long way towards that lofty aim.

Chapter 4

Optimal use of multi-informant data on co-occurrence of internalizing and externalizing problems. *

Arjen Noordhof, Albertine J. Oldehinkel, Frank C. Verhulst, Johan Ormel

Abstract

Strong between-informant discrepancies are found in ratings of (pre)adolescent problems and in co-occurrence rates between different domains of psychopathology. These discrepancies can be caused by differences in the context of measurement and the perspective of informants (Kraemer et al., 2003). The aim of this study was to develop a 'Multi-Informant Co-occurrence' model (MIC), which takes into account these differences in context and perspective. In a population-based cohort of (pre)adolescents (n=2230) from a longitudinal study in the North of the Netherlands, internalizing (INT) and externalizing (EXT) problems were rated by the (pre)adolescents themselves, their teachers, and their parents. As hypothesized Principal Component Analysis revealed four independent main components: Between-domain convergence was captured by a severity component (S), while between-domain discrepancy was captured by a direction component (D). Between-informant discrepancies were captured by a perspective (P) and a context (C) component. The use of this MIC-model will increase reliability and validity of measures of psychopathology and the four components each provide useful specific information.

Introduction

There is no gold standard for the measurement of psychiatric symptoms and disorders, and reports from different informants tend to correlate only moderately (Achenbach, et al., 1987). Therefore, in most circumstances using multiple informants seems the best strategy to chart mental health problems (Offord, et al., 1996), and there is a need for theory-based approaches to combine their reports (De Los Reyes & Kazdin, 2005). Based on a testable theory about why and how informants converge and diverge, Kraemer and colleagues (2003) have proposed a pragmatic approach for combining information from multiple sources, involving a clear and applicable research design, and a straightforward statistical procedure. This method has been applied successfully to reports of psychopathology (Kraemer, et al., 2003).

While Kraemer and colleagues (2003) applied their model to internalizing and externalizing problems separately, scores of these two problem domains typically correlate substantially (Lilienfeld, 2003). Such correlations can be related to

* *Published as an article in 2008 in The International Journal of Methods in Psychiatric Research, vol. 17(3), pg. 174-183.*

comorbidity of underlying disorders, but might be influenced by informant biases as well[6]. Co-occurrence rates tend to diverge strongly between informants (Keiley, Bates, Dodge, & Pettit, 2000), and different procedures of combining multi-informant information can lead to very different estimates of co-occurrence (Youngstrom, Findling, & Calabrese, 2003). Furthermore, Burt and colleagues (2005) showed that heritability rates of co-occurrence (of various externalizing disorders) depend on the way data from multiple informants are combined. Thus, the issue of convergence and divergence between informants is not only relevant within each domain, but also in understanding the relation between the domains of internalizing and externalizing problems. In the present study we extended Kraemer and colleagues' model (2003) to analyze the co-occurrence of problems from both the internalizing and the externalizing domain in a general population sample of pre- and early adolescents.

Central to the approach of Kraemer and colleagues (2003) is the idea that the reports of informants are not simply measurements of a well-known characteristic (G), but are influenced by the context (C) of observation and the perspective (P) of a specific informant Thus, the score of an informant is not just the result of the characteristic and measurement error, but is also influenced by context and perspective. Taking context, perspective, and measurement error into account may increase the reliability and validity of the estimated mental health problems, and provide information on the role of these aspects.

In (pre)adolescents two important contexts (C) are school and home, and two important perspectives (P) self-report (from inside) and other-report (from outside). To obtain information on these components, it is essential to carefully select informants that are likely to diverge strongly on one aspect, but converge on the other. For example, teachers and parents are likely to diverge strongly regarding the context, but both represent an 'outside'-perspective. Youngsters themselves on the other hand are likely to diverge most strongly from both their parents and teachers based on the difference in perspective. As Kraemer and colleagues (2003) substantiated in their study, the reports from teachers, parents and the (pre)adolescents themselves are well-suited to estimate these C and P components independently from the general characteristic (G). Using Principal Component Analysis (PCA) for reports of these three informants in three different samples, they found support for the hypothesis that: [1] the informants had similar coefficients on the general component (G); [2] parents and teachers had similar coefficients on the perspective component (P), both reverse to the self-report coefficients; and [3] parents and teachers had opposite coefficients on the context component (C), while the self-reports were in between.

[6] *As has been discussed by Lilienfeld (2003) and Meehl (2001), the term comorbidity is meaningful only in the context of well-validated disease entities. As still little is known about diseases underlying psychiatric classifications, we prefer using the term co-occurrence as indicating that a person can be classified in more than one psychiatric category.*

The orthogonality of these components (i.e. G, P and C do not correlate) follows from their definition. Perspective is defined as the variance that is only due to the difference between self and others. Context is defined as the variance that is only due to the difference between ratings at school and at home. As the teacher- and parent-ratings diverge on C, while they converge on P, these components are defined as orthogonal. G is defined as the convergence of all informants, which is orthogonal to C and P, because C and P are defined by divergence of two of the informants.

The aim of the present study was to extend this model, in order to capture co-occurrence of problems from the internalizing (INT) and externalizing (EXT) domain. Co-occurrences of different psychiatric categories can be captured using a hierarchical model (Krueger, 1999; Krueger, et al., 2003; Vollebergh, et al., 2001). Therefore, as shown by Weiss, Susser and Catron (1998), disorders can be described as combinations of *narrow-band* features, differentiating specific categories (e.g. depression, anxiety), *broad-band* features, differentiating internalizing and externalizing disorders, and *common* features, related to the general severity of disorders. A similar approach is to discern between a severity component (S) and a direction component (D), which correspond to the common and broad-band specific aspects respectively (Essex, et al., 2006; Ormel, et al., 2005). These S- and D-scores are by definition uncorrelated and easily computable (Essex, Klein, Cho, & Kraemer, 2003). The strength of that approach is that it captures both the co-occurrence of disorders and their differentiation. We expected that combining this approach with the model by Kraemer et al. (2003) would result in a useful tool for analyzing multi-informant measurements of co-occurring psychiatric problems.

We used the reports of different informants on a number of specific scales from both the INT and EXT domain. Instead of a different model for each informant or for each specific disorder, one model was developed that would explain the covariances between all these scores. It was hypothesized that [1] between-informant discrepancies can be explained by the C- and P-components, which should have the same patterns of coefficients as those found by Kraemer and colleagues (2003); [2] between-domain discrepancies can be explained by broad-band features, resulting in a 'direction'-component (D) on which problem-scales from the INT and EXT domain have opposite coefficients; and [3] between-domain convergence can be explained by common features, resulting in a 'severity'-component (S) on which problem scales from the INT and EXT domain have equal coefficients. Thus, to explain both between-informant and between-domain covariance, at least four components (C, P, D and S) are necessary. In principle, however, the model does not exclude additional components; because context (C) and perspective (P) might be related to common features, broad-band features, or both; and because the narrow-band features might influence the ratings as well. Thus, the analysis was aimed at testing the hypothesis that at least four components (C, P, D and S) would be found, and at exploring the possibility that more components would emerge. Furthermore, we aimed at developing a method by which these components can easily be applied in research and clinical practice.

Methods

Sample

Subjects were participants of the 'Tracking Adolescents' Individual Lives Survey' (TRAILS), a prospective cohort study of Dutch (pre)adolescents. The present study involves data from the first (T1) and second (T2) assessment wave of TRAILS, which ran from March 2001 to July 2002 and September 2003 to December 2004, respectively. A detailed description of the sampling procedure and methods is provided by De Winter and colleagues (De Winter, et al., 2005).

Briefly, the TRAILS target sample involved all 10- to 11-year-old children living in the three largest cities and some rural areas in the North of The Netherlands. Of the eligible children, 76.0% (n=2230) were enrolled in the study. Responders and non-responders did not differ regarding the prevalence of teacher-rated problem behavior and associations between sociodemographic variables and mental health indicators (de Winter et al., 2005).

Of the 2230 baseline (T1) participants, 96.4% (n=2149, 51.2% girls) participated in the first follow-up assessment (T2), which was held 2–3 years after T1 (mean number of months 29.44, S.D.=5.37). Mean age at T2 was 13.55 (S.D.=0.54). After complete description of the study to the subjects, written informed consent was obtained from both the (pre)adolescent and the parents.

Instruments

At T1 and T2, questionnaires were completed by teachers, parents and the (pre)adolescents themselves. The parent-rated Child Behavior Checklist (CBCL; Achenbach, 1991a), the Youth Self Report (YSR; Achenbach, 1991c) and the Teacher Checklist of Psychopathology (TCP; De Winter et al., 2005) were used to assess psychopathology. The CBCL and YSR have been developed for the multi-informant assessment of child and adolescent psychopathology. In these 112-item questionnaires, the informant rates descriptions of emotions and behaviors as not [0], sometimes [1], or very often [2] present. Factor analysis on these items revealed a structure of eight syndrome scales (De Groot, Koot, & Verhulst, 1994). Three of these were related to the internalizing domain (INT): 'Anxious-depressed' (Anx), 'Somatic complaints' (Sc), and 'Withdrawn/depressed' (Wd). Two were related to the externalizing domain (EXT): 'Aggressive behavior' (Agg) and 'Rule-Breaking behavior' (Rb). The TCP contains descriptions (vignettes) of problem behaviors, corresponding to the eight syndromes of the CBCL and YSR. Teachers rated each of these vignettes on a 5-point rating scale. At T1, for 79% of the participants these questionnaires were completed by all three informants, and for 97% by at least two informants. At T2, for 60% of the original (T1) participants, questionnaires were completed by all informants, and for 91% by at least two informants. The low three-informant rates at T2 were mainly caused by a relatively low response rate among teachers. Only children for which three informants completed the questionnaires were included in the analyses.

Statistical analysis

All analyses were performed using *SPSS 12.0.2*. For each of the five syndrome scales related to the INT and EXT domain, we calculated a mean item score (i.e. five scores for each of the three informants). Principal Component Analysis (PCA) with these 15 (3x5) subscales was used to test the hypotheses. As explained in the introduction the components are defined as orthogonal and are hypothesized to explain a large part of the variance in the subscales. For these specific hypotheses Principal Component Analysis (PCA) is an appropriate tool, as it is a method that maximizes the explained variance of each orthogonal component.

Following the approach of Kraemer and colleagues (2003) we used the unrotated principal components. Subsequently, the stability of the model across different ages was investigated, by performing the same analyses on the T2-data.

Table 1. *Contrasts used for the interpretation of the predicted PCA-components.*

	Interpretation of coefficients				
Severity	General (**G**)				
Direction	Internalizing (**I**)	*vs.*	Externalizing (**E**)		
Perspective	Self (**Se**)	*vs.*	Others (**Ot**)		
Context	Home (**Ho**)	*vs.*	School (**Sc**)	*vs.*	Both (**B**)

Table 1 shows the four different components that we expected based on these hypotheses, and for each of these components the expected contrasts. For all components we evaluated whether they matched the hypothesized pattern of: either [1] a severity component (S), indicated by similar coefficients of the subscales (table 1: G); [2] a direction component (D), indicated by a contrast between coefficients of INT and EXT subscales (table 1:I vs E); [3] a perspective component (P), indicated by a contrast between self-rated and other-rated problems (table 1: Se vs Ot); or [4] a context component (C), indicated by a contrast between school and home (table 1: Ho vs Sc, and B for involvement in both contexts). All components that could be interpreted this way were included in the final model. Other components that were found in the analyses could not be interpreted a-priori. Such components can be either related to unique variance of a specific subscale, or to domain- or disorder-specific informant discrepancies. The number of possibly important components is high and post-hoc interpretations of these components can be rather arbitrary. To prevent irrelevant chance findings, other components were only interpreted (post-hoc) if their eigenvalue exceeded 1. The rationale for this, is that if the eigenvalue of a component does not exceed 1, that component contributes less explained variance than a single subscale in the original model (i.e., less than 7%). Variance that cannot be explained by

our model is to be explained by specific subscales and specific biases that are certainly of interest, but require more specific hypotheses, which are beyond the scope of this article.

In order to enhance the applicability of the model for research and clinical practice we developed algorithms for S, D, P and C that can be made without performing a PCA. We followed the method used by Essex et al. (2003) to compute the S- and D-component, and extended it to the computation of the P- and C-component. We investigated to what extent individual scores based on these algorithms were comparable to scores based on the coefficients of the principal components by calculating individual component scores with the SCORES-option available in the SPSS factor analysis command (which can be used for PCA as well) and correlating these scores with the scores based on the simple algorithms. The severity component (S) was calculated by taking the mean of all 15 subscores; the direction component (D) by contrasting the internalizing and externalizing subscales; the perspective component (P) by contrasting ratings by others and self ratings; and the context component (C) by contrasting teacher and parent ratings. More specifically:

[1] S = mean (all subscores)
[2] D= [mean (internalizing subscales) – mean (externalizing subscales)] / 2
[3] P = [mean (parent and teacher subscales) – mean (self subscales)] / 2
[4] C= [mean (teacher subscales) – mean (parent subscales)] / 2

Results

Analyses were performed on data from the first (T1) and second (T2) assessment. Means and standard deviations of these data are reported in table 2.

Table 2. *Means and standard deviations for YSR, CBCL and TCP subscales at the first and second assessment wave.*

Informant	Subscale	data T1		data T2	
		Mean	*Sd*	Mean	*Sd*
Self	Anx	0.33	*0.27*	0.31	*0.29*
	Sc	0.43	*0.31*	0.33	*0.29*
	Wd	0.34	*0.29*	0.34	*0.30*
	Agg	0.31	*0.25*	0.31	*0.24*
	Rb	0.22	*0.17*	0.26	*0.20*
Parent	Anx	0.28	*0.25*	0.20	*0.22*
	Sc	0.20	*0.21*	0.16	*0.20*
	Wd	0.25	*0.27*	0.24	*0.28*
	Agg	0.35	*0.29*	0.23	*0.25*
	Rb	0.13	*0.13*	0.10	*0.14*
Teacher	Anx	0.69	*0.95*	0.78	*1.03*
	Sc	0.58	*0.86*	0.70	*0.96*
	Wd	0.72	*0.99*	0.98	*1.14*
	Agg	0.62	*0.98*	0.65	*1.10*
	Rb	0.27	*0.71*	0.32	*0.84*

Note: *Anx* = Anxiety; *Sc* = Somatic complaints; *Wd* = Withdrawn behavior; *Agg* = Aggression; *Rb* = Rule-Breaking behavior.

Analysis of T1-data

The result of the PCA are reported in table 3. Four components were found that had eigenvalues greater than 1 and followed the hypothesized between-informant and between-domain convergences and divergences. Other components did not meet the criterion of an eigenvalue greater than 1, and therefore were not included in the final model. Therefore, four components were included in the final model, which explained 64% of the total variance.

Two components were found with a rather high between-informant convergence. These were interpreted as a severity component (S), with positive coefficients for all subscales, and a direction component (D), with a contrast between the coefficients of the INT (+; positive coefficients) and the EXT (-; negative coefficients) domain. The other two components matched the hypothesized between-informant divergences: a perspective component (P), with a strong contrast between the coefficients of the self-rated (-) and the other-rated (+) subscales; and a context component (C), with a strong contrast between the coefficients of the parent-rated (-) and the teacher-rated subscales (+), and with the self-rated subscales in between (only weak coefficients <.20).

Table 3. *Component coefficients for the joint analysis of the internalizing and externalizing domain.*

Informant	Subscale	**S**everity		**P**erspective		**D**irection		**C**ontext	
	Anx	0.58	**G**	-0.51	**Se**	0.30	**I**	0.16	**B**
	Sc	0.46	**G**	-0.49	**Se**	0.15	**I**	0.15	**B**
Self	Wd	0.58	**G**	-0.45	**Se**	0.28	**I**	0.19	**B**
	Agg	0.62	**G**	-0.51	**Se**	-0.26	**E**	0.06	**B**
	Rb	0.52	**G**	-0.47	**Se**	-0.35	**E**	0.05	**B**
	Anx	0.59	**G**	0.23	**Ot**	0.35	**I**	-0.41	**Ho**
	Sc	0.44	**G**	0.12	**Ot**	0.30	**I**	-0.34	**Ho**
Parent	Wd	0.58	**G**	0.28	**Ot**	0.31	**I**	-0.32	**Ho**
	Agg	0.67	**G**	0.27	**Ot**	-0.24	**E**	-0.42	**Ho**
	Rb	0.61	**G**	0.25	**Ot**	-0.33	**E**	-0.41	**Ho**
	Anx	0.42	**G**	0.45	**Ot**	0.22	**I**	0.50	**Sc**
	Sc	0.37	**G**	0.35	**Ot**	0.20	**I**	0.48	**Sc**
Teacher	Wd	0.32	**G**	0.43	**Ot**	0.33	**I**	0.47	**Sc**
	Agg	0.43	**G**	0.27	**Ot**	-0.61	**E**	0.30	**Sc**
	Rb	0.37	**G**	0.24	**Ot**	-0.61	**E**	0.27	**Sc**
% Explained variance		**27%**		**14%**		**12%**		**11%**	

Note: *Anx* = Anxiety; *Sc* = Somatic complaints; *Wd* = Withdrawn behavior; *Agg* = Aggression; *Rb* = Rule-Breaking behavior, *G*=general; *Se*=self; *Ot*=other; *I*=internalizing; *E*=externalizing; *Ho*=Home; *Sc*=School; *B*=both.

Replication at T2

Results of the PCA on T2-data are shown in table 4. The same components were found, with only small differences in component-coefficients. One exception was that the two teacher-rated externalizing subscales (Agg and Rb) loaded less strongly on the P-component than at T1. Only minor changes in explained variance of the four components occurred. This model explained 64% of variance, which is exactly the same as the model based on T1-data. The solution for the D-component was opposite to the solution found with the T1-data, as the externalizing scales loaded positive and the internalizing scales loaded negative. However by multiplying all scores with -1, which is statistically justified, it is possible to obtain a completely similar solution, that is with high D-scores indicating relatively more internalizing problems.

Table 4. *Component coefficients for the joint analysis of the internalizing and externalizing domain at T2.*

Informant	Subscale	**S**everity		**D**irection		**P**erspective		**C**ontext	
	Anx	0.54	**G**[2]	-0.49	**I**	-0.46	**Se**	0.09	**B**
	Sc	0.48	**G**	-0.32	**I**	-0.45	**Se**	0.07	**B**
Self	Wd	0.52	**G**	-0.47	**I**	-0.35	**Se**	0.10	**B**
	Agg	0.60	**G**	0.17	**E**	-0.54	**Se**	0.06	**B**
	Rb	0.53	**G**	0.34	**E**	-0.45	**Se**	0.10	**B**
	Anx	0.63	**G**	-0.21	**I**	0.30	**Ot**	-0.39	**Ho**
	Sc	0.52	**G**	-0.19	**I**	0.11	**Ot**	-0.26	**Ho**
Parent	Wd	0.59	**G**	-0.11	**I**	0.38	**Ot**	-0.37	**Ho**
	Agg	0.68	**G**	0.32	**E**	0.21	**Ot**	-0.37	**Ho**
	Rb	0.65	**G**	0.46	**E**	0.12	**Ot**	-0.30	**Ho**
	Anx	0.46	**G**	-0.23	**I**	0.41	**Ot**	0.53	**Sc**
	Sc	0.43	**G**	-0.18	**I**	0.33	**Ot**	0.48	**Sc**
Teacher	Wd	0.31	**G**	-0.33	**I**	0.53	**Ot**	0.36	**Sc**
	Agg	0.35	**G**	0.66	**E**	0.05	**Ot**	0.38	**Sc**
	Rb	0.35	**G**	0.62	**E**	0.01	**Ot**	0.37	**Sc**
% Explained variance		**27%**		**14%**		**13%**		**10%**	

Note: *Anx* = Anxiety; *Sc* = Somatic complaints; *Wd* = Withdrawn behavior; *Agg* = Aggression; *Rb* = Rule-Breaking behavior, *G*=general; *Se*=self; *Ot*=other; *I*=internalizing; *E*=externalizing; *Ho*=Home; *Sc*=School; *B*=both.

Correspondence to easily computable scores

As shown in table 5, correlations between the algorithms for S, P, D and C and the component-scores derived from the PCA-model were all well above .90. This result was found with the T1-data as well as T2-data, indicating that these algorithms represent the component scores of the PCA model almost perfectly.

Table 5. *Correlations, based on T1-data and T2-data, between the PCA-components and the algorithms for Severity, Direction, Perspective and Context.*

PCA	Algorithm	T1	T2
Severity	mean (all subscores)	*0.99*	*0.99*
Direction	[mean (internalizing subscales) – mean (externalizing subscales)] / 2	*0.98*	*0.94*
Perspective	[mean (parent and teacher subscales) – mean (self subscales)] / 2	*0.96*	*0.93*
Context	[mean (teacher subscales) – mean (parent subscales)] / 2	*0.93*	*0.94*

Discussion

This study extended Kraemer and colleagues'(2003) multi-informant approach by modeling adolescents' internalizing and externalizing problems and their co-occurrence. We found support for a four component model of between-informant and between-domain convergences and divergences. The results support the hypotheses of this study: [1] between-informant discrepancies can be captured by a context (C) and a perspective (P) component; [2] between-domain discrepancies can be captured by a direction component (D); and [3] between-domain convergence can be captured by a severity component (S). This 'Multi-Informant Co-occurrence' model (MIC) was replicated with follow-up data from the same sample. The PCA-components almost perfectly correlate with the algorithms we developed for S, P, D and C. The formulas for these scores were directly derived from the hypotheses of this study. Therefore, this result can be interpreted as strong support for our main hypotheses. Provided that these results can be replicated in different samples, these algorithms allow for application of the model in research with small samples and individual diagnosis. While supporting the main hypotheses of this study, these results do not imply that the original model can be *replaced* by 4 Principal Components. We chose not to interpret more than 4 components, for reasons explained in the methods section. It was not expected that all variance could be explained by the MIC-model, as this is a model on broad-band specific features. To explain more of the variance the model would have to be extended to cover narrow-band specific features and more specific informant-discrepancies.

Using the MIC-model will increase both the validity and reliability of the measurement of emotional and behavioral problems. The reliability increases, because combining informants reduces measurement error. An extra advantage of aggregating information from multiple sources is that variance in the estimation of marginal parameters (e.g. prevalence, incidence) and associations (e.g. in regression analysis) is reduced (Hofler, 2005). Most importantly the validity increases by explicitly including the context (C) and perspective (P) in a multiple disorder model. Formulating and testing hypotheses regarding the reasons for discrepancies between specific informants is a more valid method for dealing with informant-discrepancies than only taking a mean score. Furthermore, the four-component MIC-model allows to take into account the high co-occurrence between the internalizing and externalizing domain by separating components with between-domain convergence (S, P and C) from a component with between-domain divergence (D).

The four components are not unidimensional scales of a specific trait, but should be interpreted in conjunction and as relative measures. The S-component can be used to discriminate between (pre)adolescents with regard to the amount of psychopathological problems. The D-component discriminates between those whose problems are relatively more internalizing and those whose problems are more externalizing. The P-component discriminates between those whose problems are relatively more self-reported, and those whose problems are more other-reported.

Finally, the C-component discriminates between those whose problems are relatively more observed at school, and those whose problems are more observed at home.

All four components contribute specific information regarding disorders and their co-occurrence. Although the relevance of each component may depend on the research question at hand, total disregard of some of these components implies using a reduced model. Although the relevance of each component may depend on the research question at hand, disregard of some of these components implies using a reduced model. Mean scores for INT and EXT, which are very often used in current research, can be transformed to S and D components without losing any information, using the provided algorithms. Using only INT and EXT mean scores , therefore implies a disregard of the P and C components.

The usefulness of the S- and D-components can be illustrated by the different ways in which temperament traits can influence the development of psychopathology (Shiner & Caspi, 2003). Such influences can be general, disorder-specific, or pathoplastic (Oldehinkel, Hartman, De Winter, Veenstra, & Ormel, 2004; Ormel, et al., 2005). The S- and D-component can be used to understand this issue of specificity better. General factors (e.g. frustration) are associated with the development of psychopathology, but not with the nature of problems. Thus, general factors should be positively correlated to the S-component, but uncorrelated to the D-component. Disorder-specific factors are associated with the development of a specific type of psychopathology, and therefore should not only be correlated to the S-component, but also to the D-component. Effortful control, for example, has been linked to the development of more externalizing problems in particular, and should therefore be positively correlated with the D-component, as a low D-score indicates relatively more externalizing problems. Pathoplastic factors are only associated with the nature, but not the amount, of problems, so these should only be correlated with the D-component. For example, shyness has been suggested as a component that is associated with relatively more internalizing problems, but not with the amount of problems. Accordingly, it should be positively correlated with the D-component, and uncorrelated with the S-component. An example of the application of the S- and D-components in the study of risk-factors can be found in Essex et al. (2006).

The P- and C-components contain additional information regarding the problems of (pre)adolescents. This information can, for example, be used to assess whether school-related factors (e.g. bad marks, truancy) are specifically related to problems at school, and therefore only associated with the C-component, or to more generalized problems, and therefore only associated with the S-component. Another example is the possibility to assess whether the P-component is related to the amount of psychosocial support a (pre)adolescent receives. A low score on the P-component indicates that problems are relatively more self-reported. These problems may remain untreated as the social environment does not recognize them as such. An example of the use of a context (C) and perspective (P) component is given by Perren et al. (2006) who show that differences in perspective between children and adults in hyperactivity/ impulsivity predict peer rejection.

The results provide rather strong support for the MIC-model. A strength of the study is, that it is based on a large cohort from the general population with limited non-response. A drawback is that vignettes were used to assess teacher ratings, instead of the original Teacher Rating Form (TRF; Achenbach, 1991b). The TRF has the advantage of using nearly the same items as the CBCL and YSR, so there might be less bias due to differences in the measurement-instrument used. However, the TCP-items follow the TRF-scales, and correlations between these measurements (Ferdinand, R.F., personal communication) indicate that the TCP-vignettes can be used as a proxy for the TRF scales. In the present study, we replicated the findings using follow-up data of the same sample, but replication in other samples is needed as well, among other things to assess the generalizability to other age groups and clinical populations.

The approach proposed in this article is preferable to a separate analysis of the subjective reports of different informants, because combining different perspectives and contexts results in a more objective measure of psychopathology. The approach is also preferable to separate analysis of the domains of internalizing and externalizing problems, because it allows to distinguish between common and broadband-specific effects. These are clear advantages for any research on psychopathology and especially for research on co-occurrence. The model could be used in clinical practice as all four components give important information regarding the condition of a patient which may be overlooked when looking to specific informant reports on specific disorders. It offers a clear method of combining different kinds of information that seems preferable to the situation where clinicians read all reports separately and combine this information subjectively. The algorithms we developed (see table 5) allow for easy, paper-and-pencil, computation of the S, P, D and C components. However, it is important to underscore that the interpretation of such scores at an individual level may not be straightforward (De Los Reyes & Kazdin, 2005), because some informant-discrepancies may have individual specific reasons.

As a final remark we would like to emphasize that the MIC-model is work in progress. The useful advantages and applications we have described are dependent on proper replications in other (e.g. clinical) samples. There are various possibilities for further development of the model. Firstly, the method is not specific to the instruments we used, nor to common mental health problems, nor to youngsters. Therefore, different measurement instruments can be used and the method can be expanded to other problem domains (e.g. ADHD symptoms, psychotic experiences, substance abuse). Secondly, using additional informants, for instance peers, would provide more information regarding different contexts or perspectives. Thirdly, future research may challenge some of the assumptions of the model. Context and perspective are defined as orthogonal components, but the way context and perspective shape an impression of someone and how that impression translates into questionnaire responses may not be fully captured by independent components, and deserves further investigation. Nevertheless, we expect this new tool to be useful for multiple purposes and encourage its application and further development.

Chapter 5

Comorbidity between Internalizing and Externalizing problems in adolescence: fact or artefact?

Arjen Noordhof, Albertine J. Oldehinkel, Johan Ormel

Abstract

Substantial correlation has been found between the domains of Internalizing and Externalizing problems (i.e. IE-correlation). Several models have been proposed to explain IE-correlation as the result of causal relations between the problems of these domains or shared risk factors. An alternative explanation is that the correlation results from method bias. Five method biases are presented that may result in an overestimation of IE-correlation: attrition bias, Berksonian bias, stratification by age and gender, observation bias, and response bias. In the research presented here it was tested to what extent these biases may explain the observed correlation between Internalizing and Externalizing problems. The hypothesis that all IE-correlation results from methodological artefacts could not be rejected. The estimation of IE-correlation appeared highly sensitive to the specific informants and instruments that are used.

Introduction

The domains of Internalizing and Externalizing problems show considerable correlation (Krueger & Markon, 2006a). This may be explained by causal models that link disorders or problems from these domains (Neale & Kendler, 1995), but correlations may also be caused by methodological biases (Lilienfeld, 2003). As shown by Hofler, Lieb, & Wittchen (2007) such biases can have an important impact on estimates of the association between two disorders. If the correlation between Internalizing and Externalizing scales can be explained by biases, then substantive explanations of comorbidity may actually be explanations of an artefact. Therefore, testing the hypothesis that the correlation between Internalizing and Externalizing problems is caused by methodological bias is crucial for understanding the structure of psychopathology.

Several methodological biases can influence the correlation between scales. In this study we will investigate sampling bias and measurement bias. Sampling bias results from the fact that correlations are estimated in a specific sample. Important sampling biases are attrition bias, Berksonian bias, and bias due to population stratification. Measurement bias results from the fact that correlations are estimated on the basis of subjective reports on specific instruments. Measurement bias may involve observation bias or response bias. Sampling bias and measurement bias can result in discrepancies between samples, instruments and informants regarding the estimated correlation between Internalizing and Externalizing problems (in the following: IE-correlation).

However, discrepancies are not necessarily caused by methodological bias, which will be explained in the following sections and is illustrated in figure 1.

Attrition bias

Attrition bias results from the selective loss of participants from a sample. Subjects with comorbid conditions may drop out of a study more easily because of their problems. Alternatively, subjects with only Externalizing problems may drop out more easily. Attrition can result in underestimation or overestimation of IE-correlation in the incomplete dataset.

Berksonian bias

Berksonian bias refers to the possibility that chances to receive treatment are higher for people with comorbid conditions. This can result in an overestimation of the correlation between disorders in a clinical sample. However, differences between clinical and population samples are not necessarily due to Berksonian bias: qualitative differences between patients and non-patients in the mechanisms that cause comorbidity may create discrepancies between samples as well.

Population stratification

Population stratification (also referred to as ecological fallacy or Simpson's paradox) refers to the phenomenon that if a sample consists of homogeneous subgroups, associations in the total sample are reflecting both within-subgroup and between-subgroup associations. Kraemer, Wilson, & Hayward (2006) showed that if a sample consists of different age groups, and the prevalences of two psychiatric disorders increase with age, a correlation between the disorders can be found even if the disorders are uncorrelated at each age. Population stratification may involve all kinds of subgroups in a population, but in the current study we focus on age and gender. Bias due to population stratification would result in a discrepancy between the association found in these subgroups and the association found in the total population.

Observation bias

Reports of different informants are based on the observations that these informants make. These observations may be accurate, but may also be biased. Observed problems in one domain may bias a rater towards observing more problems in another domain as well (Lilienfeld, 2003). Specific mental health problems in children may lead to conflicts with their parents, which in turn may lead to a general negative parental attitude and negative observations in multiple domains, hence to a higher correlation between problem domains. If children are demoralized and pessimistic due to problems they experience in a specific domain, they may become increasingly aware of problems in other domains as well. Almost every child sometimes feels depressed, anxious, or angry. If these normative behaviors are interpreted and reported as problematic because of problems in another domain, the domains will correlate even if underlying disorders are not causally related in any other way than via observation.

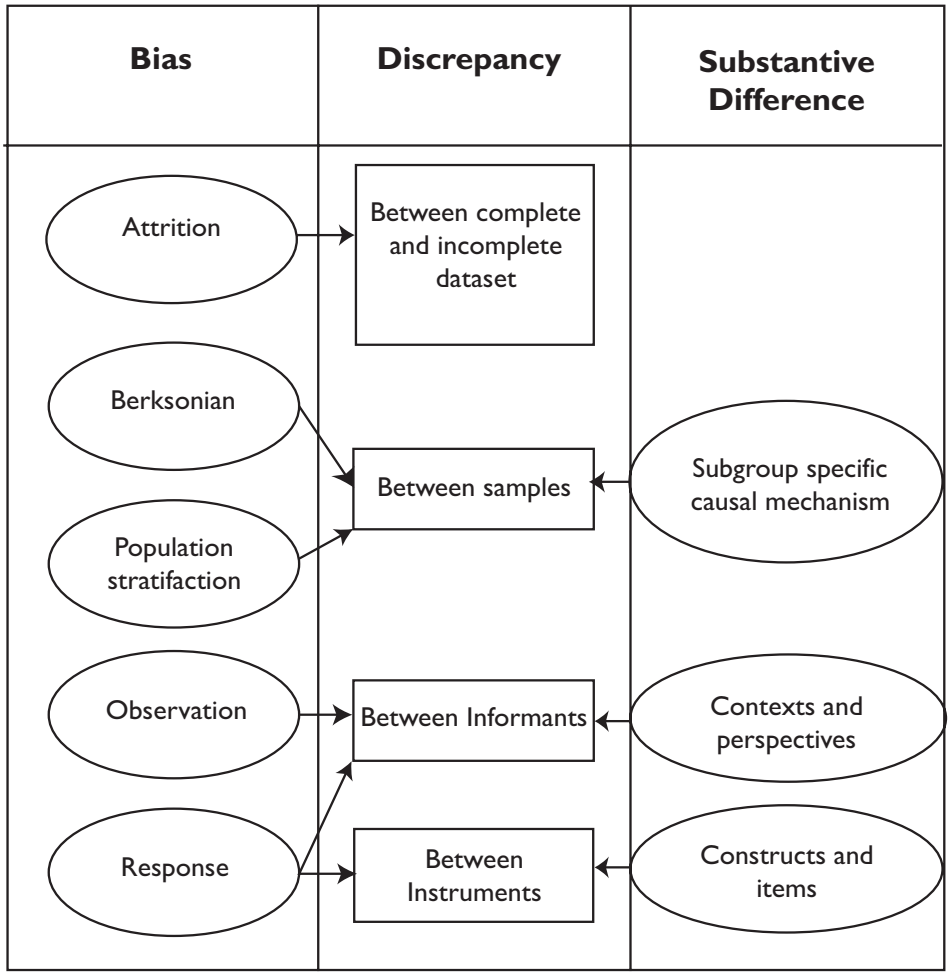

Figure 1. Causes of discrepancy between samples, informants, and instruments.

Observation biases of different informants may be related. For example, they may be influenced by interactions between child and parent or between parent and teacher. On the other hand, it is improbable that biases are identical, so observation bias will usually cause informant discrepancy. Such discrepancy can also be caused by actual differences in behaviors and emotions that can be observed. Informants make their observations in different contexts (e.g., at school or at home), and there is a difference between what can be observed by persons themselves and what can be observed by others (see Kraemer, et al., 2003; Noordhof, et al., chapter 4 of this thesis). Furthermore, the unique view of a single informant is not necessarily wrong. If children report a lot of problems this should be regarded as an indication that they are suffering and not be disregarded as 'bias'. Finally, not all observation biases result in an overestimated IE-correlation. A teacher may have a tendency to contrast children in a classroom and thereby increase the difference between Internalizing and Externalizing problems ('contrast bias'). Externalizing problems may attract so much attention that co-occurring Internalizing problems remain unrecognized. Altogether, observation biases may cause informant discrepancies, but discrepancies do not necessarily reflect a biased IE-correlation.

Response bias

Responses on a measurement instrument are not only influenced by subjective observation, but also by the design of the instrument and the way it is used by an informant. For example, some people have a tendency to agree with questionnaire items regardless of their content (acquiescence). If an instrument has a three point rating scale (0,1,2), some people have a tendency to give moderate responses when they are uncertain (many 1-scores), while others have a more extreme response style (many 0- or 2-scores). De Jonge & Slaets (2005) have shown that such response biases emerged even in questionnaires without content (i.e. only responses, no questions), and were correlated with personality traits of the responders.

Specific response styles of informants can be triggered by the design of an instrument and may result in biased IE-correlation. Response style is a characteristic of an informant and it is unlikely that biases of different informants are strongly correlated. Therefore, response bias is likely to result in informant discrepancy. Also, instruments may differ in the kind and amount of response bias that they trigger, because of differences in design like item order, response format and explanatory text. However, instruments may also differ in item content. Instrument discrepancy can therefore be caused by response bias, but also by substantive differences between instruments.

Detecting and exploring discrepancies

In the current study we examined to what extent the five above-mentioned biases contribute to IE-correlation. To this end we compared reports of Internalizing and Externalizing problems in multiple samples and on the basis of multiple informants and

instruments. As illustrated in figure 1, discrepancies between samples, informants, and instruments may result from biases, but also from substantive differences. On the basis of this figure one can argue that, in the absence of a 'gold-standard', measurement bias can only be defined if the bias itself is known or if all other causal influences are known. We do not deal with such ideal situations and therefore discrepancies have to be interpreted on the basis of hypotheses regarding both biases and substantive differences.

With regard to differences between samples we did not have a priori hypotheses. With regard to informant discrepancies, we tested to what extent IE-correlation can be explained by only assuming within-informant IE-correlation, i.e. correlation between subscales of reports by the same informant. IE-correlation that can not be fully explained by within-informant correlation supports the existence of actually co-occurring Internalizing and Externalizing problems, which can be observed by multiple informants. To test this hypothesis we employed Confirmatory Factor Analysis for multi-trait multi-method modelling (CFA-MTMM), which will be explained in more detail in the Methods section. These models allow to distinguish between informant specific IE-correlation and IE-correlation that results from covariance between multiple informants. As will be further discussed in the Methods section, none of these models provides a fully satisfactory distinction between actual IE-correlation and bias. While some MTMM models will probably result in overcontrolling for bias, others result in factors that may still contain observation and response bias. As we did not find or develop a better solution, we will report estimations of IE-correlation on the basis of state-of-the-art MTMM models, and develop hypotheses with regard to informant discrepancy by comparing their differences. With regard to instrument discrepancy we also used a MTMM model in order to test to what extent IE-correlation could be attributed to within-instrument correlation, i.e. correlation between the subscales of the same instrument.

In short, we tested the hypothesis that the correlation between Internalizing and Externalizing problems is caused by method biases rather than the latent structure of psychopathology by comparing different samples, informants and instruments.

Methods

Sample

Subjects were participants in the 'Tracking Adolescents' Individual Lives Survey' (TRAILS), a prospective multi cohort study of Dutch (pre)adolescents. The study involved a representative sample from the general population and is described in detail in Huisman et al.(2008).

Briefly, the target sample involved all 10- to 12-year-old children living in the three largest cities and some rural areas in the North of The Netherlands. Of the eligible children, 76.0% (n=2230, mean age = 11.09, SD =0.55) were enrolled in the study. Responders and non-responders did not differ regarding the prevalence of teacher

rated problem behavior and associations between sociodemographic variables and mental health indicators (De Winter, et al., 2005). To date, the population cohort has been assessed three times (T1: March 2000- July 2001, T2: September 2003-December 2004, T3: September 2005-December 2007). Participation rates were 96.4% at T2 (mean age= 13.55, SD = 0.53), and 81.4% at T3 (mean age= 16.25, SD = 0.73). After complete description of the study to the subjects, written informed consent was obtained from the parents at each assessment wave and from the adolescents at T2 and T3.

The clinical cohort target sample involved children who had been referred to a child psychiatric outpatient clinic in the Northern Netherlands at any point in their life. Of the eligible children 43.0% (n=543) were enrolled in the study. Responders and non-responders did not differ regarding the prevalence of teacher rated problem behavior (Huisman, et al., 2008). We used data from the first assessment wave, which ran from September 2004 to December 2005 (mean age = 11.11, SD =0.50). After complete description of the study to the subjects, written informed consent was obtained from the parents.

Instruments

• *CBCL, YSR and TRF*

In both cohorts, the parent rated Child Behavior Checklist (CBCL; Achenbach, 1991a; Verhulst, et al., 1996) and the Youth Self Report (YSR; Achenbach, 1991c; Verhulst, van der Ende, & Koot, 1997b) were used to assess psychopathology at all measurement waves. In the clinical cohort the Teacher Report Form (TRF; Achenbach, 1991b; Verhulst, van der Ende, & Koot, 1997a) was also used. In the TRAILS-study the CBCL was completed by one of the parents, which was the mother in most cases. The CBCL, YSR, and TRF are 112-item questionnaires in which informants rate descriptions of emotions and behaviors on a 3-point scale (not [0], sometimes [1], or very often [2]). The period over which they are asked to report is the last six months. Factor analysis on these items has revealed a structure of eight syndrome scale (Achenbach, 1991a-c). Three of the scales are related to the Internalizing domain (INT): Anxious-Depressed (Anx, 13 items, α=0.78), Somatic complaints (Sc, 11 items, α=0.69), and Withdrawn-Depressed (Wd, 8 items, α=0.71). Two are related to the Externalizing domain (EXT): Aggressive Behavior (Agg, 18 items, α=0.88) and Rule-Breaking behavior (Rb, 17 items, α=0.68). The other three scales were not used in the current study. In a study by Hartman et al. (1999) the distinction between an INT and EXT factor was replicated quite well for both the TRF and CBCL, although they found no significant difference in model fit between a 2-factor and an 8-factor solution.

• *RCADS*

The Revised Child Anxiety and Depression Scale (RCADS; Chorpita, Yim, Moffitt, Umemoto, & Francis, 2000) is a self report questionnaire with 47 items, which are

scored on a 4-point scale (never [1], sometimes [2], often [3], or always [4]). The questionnaire covers six subscales all related to DSM-IV syndromes from the Internalizing domain: Generalized Anxiety Disorder (6 items, á=0.78), Social Phobia (9 items, á = 0.78), Separation Anxiety Disorder(7 items, á = 0.66), Panic Disorder (9 items, á = 0.75), Obsessive-Compulsive Disorder(6 items, á = 0.68), and Major Depression Disorder(10 items, á = 0.71). The period over which these items must be rated was not specified.

• *ASBQ*

The ASBQ is comparable to the Self-Report Delinquency Scale (Moffitt & Silva, 1988), and consists of 31 items on lifetime antisocial behaviors. Respondents rate the frequency of specific antisocial behaviors on a 5-point scale (no, never [1], once [2], two or three times [3], four to six times [4], seven times or more [5]). For the current study the total score was used as a measure of Externalizing problems (á=0.88). Given the extreme skewness of this scale, we used the natural logarithm of the scores.

• *Mental care utilization*

Parents completed a questionnaire on care utilization by their child. For the current study we used the items that asked whether the child received any professional mental health care (inpatient or outpatient treatment) in the year before measurement.

Statistical analysis

All analyses were done using MPlus version 5.2. The self report (YSR) based model in the general population sample at T1 was used as a 'reference model'. That is, discrepancies are expressed as a comparison with this reference. In the absence of a gold-standard measure the choice for a reference model is inherently arbitrary. The choice for T1 was based on the fact that it is the baseline measurement in our study. The choice for self report was based on the idea that self judgement is a plausible starting point for diagnosis. The model corresponding to this reference is shown in figure 2. The factor loadings of the subscales on the Internalizing and Externalizing factor were fixed in subsequent analyses, while the correlation between these factors (in the following IE-correlation) was freely estimated. This was done to allow for a direct comparison between IE-correlations without considering differences in factor loadings between models.

All models were evaluated with the fit indices Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI). If these indices clearly indicated inadequate model fit, we exploratively improved model fit by loosening some constraints. For example, by freely estimating factor loadings or by allowing some residual correlations. In those cases we report IE-correlations for both the original and the exploratively improved models.

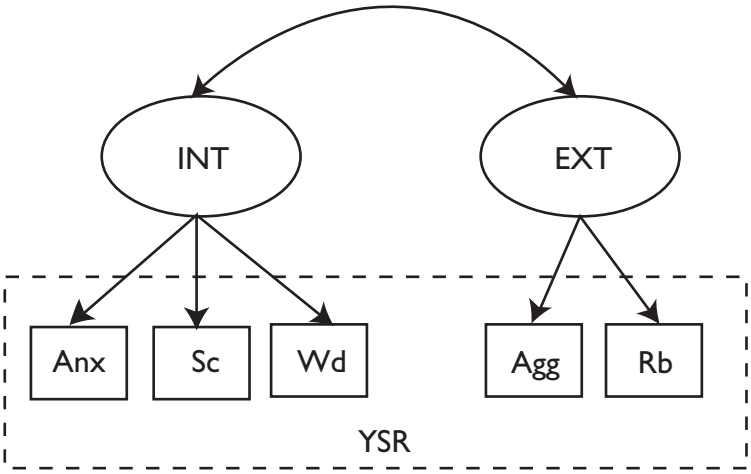

Figure 2. Reference model of correlated Internalizing and Externalizing problems.
Note: Anx = Anxious-Depressed; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed, *Agg* = Aggressive Behavior, *Rb* = Rule-Breaking Behavior, *INT* = Internalizing; *EXT* = Externalizing; *YSR* = Youth Self Report.

• *Sample discrepancies*

Attrition bias was investigated on the basis of parent reported (CBCL) problems from the general population T3-data, because these show the highest amount of missing data (38.4%) and therefore could be expected to be most affected by attrition bias. Missing data were imputed using the multiple imputation method NORM (NORM, version 2.03, Schafer, 2000) to create 30 imputed datasets, which were subsequently analysed with MPlus 5.2. To impute missing data, we used 88 variables of which we expected that they might influence attrition, including demographic characteristics, temperament, social skills, family functioning, IQ, parental psychopathology and child psychopathology; measured at various measurement waves (T1, T2, T3). Data were imputed in 30 iterations, based on linear regressions of these variables. We assumed missingness at random (MAR), which means that attrition results only from chance and from the modelled influence of observed variables. Models based on the basis of this extensive imputation procedure were compared with models based on listwise deletion (i.e. just deleting those cases with one or more missing values).

To investigate Berksonian bias, the reference model (figure 2) was fitted on data of the clinical cohort at T1. Furthermore, we fitted the model in a subgroup who received mental health care during the year before the questionnaire was completed (N=105, 4.7%).

To investigate population stratification, the general population sample was divided into six subgroups on the basis of age and gender. Participants were measured at different ages, but placed in only one of the subgroups to avoid dependent observations. In other words, for each subject we used the score of either T1, T2 or T3, based on random selection. The reference model (figure 2) was fitted for the aggregate sample of all subgroups and within each of the subgroups separately. If population stratification by age and gender resulted in bias, the IE-correlation would be lower in the subgroups than in the aggregate sample.

• *Informant discrepancies*

To investigate informant discrepancies we fitted Multi-Trait Multi-Method CFA models (MTMM-CFA) combining reports of all three informants. For these analyses we used T1-data from the clinical cohort, because for this sample we had information for all three informants(CBCL, YSR, TRF). Two state-of-the-art and regularly used MTMM-CFA models are the Correlated Traits-Correlated Uniqueness model with multiple indicators (CT-CU; Marsh & Byrne, 1993) and the Correlated Traits-Correlated Methods minus one model (CT-CM-1; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). As will be explained below, the CT-CU model probably overestimates bias, but the CT-C(M-1) model will not result in bias free estimations. We chose to fit both the CT-CU model and three CT-C(M-1) models in order to estimate multiple IE-correlations that range from probably overcorrected to probably biased estimations. While clearly not ideal, we considered this the most adequate way of representing informant discrepancies given the currently available methods.

In the multiple indicator CT-CU model, illustrated in figure 3, an INT and EXT factor are estimated for each informant. These factors are freely correlated within each informant (i.e. correlated uniqueness), but uncorrelated between informants. The informant specific INT and EXT factors load on cross informant, higher order, INT and EXT factor respectively. The result of this way of modelling is that IE-correlation that is unique to one informant will result in an increased correlation between informant specific INT and EXT factors. IE-correlation that it is found across multiple informants, will result in a correlation between the cross informant INT and EXT factors. This model results in overestimating bias as all informant specificity is interpreted as method bias. As has been explained in the introduction discrepancies between informants cannot be equated to method bias, because they may result from objective differences in the perspectives of informants and the contexts in which they observe the child. If not all informant specific IE-correlation can be interpreted as bias, then the correlation between the higher-order INT and EXT factors in the CT-CU model cannot be regarded as an accurate estimation of IE-correlation. Rather it should be regarded as the minimum amount of cross-informant IE-correlation that has to be assumed on the basis of reports of three informants.

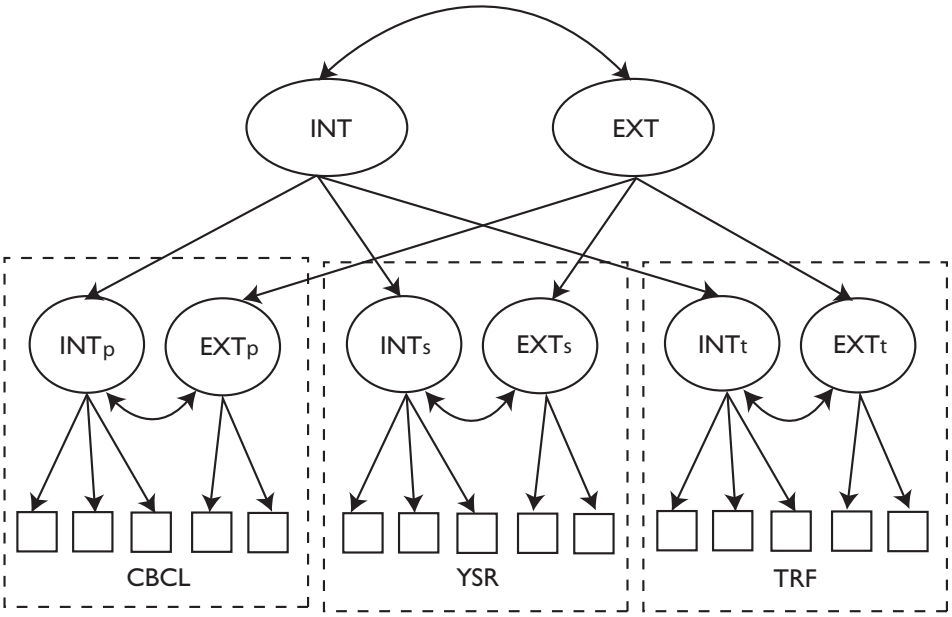

Figure 3. Multiple Indicator Correlated traits – Correlated Uniqueness model with three informants.

Note: The open squares at the bottom of the figure refer to the subscales of the instruments used and are shown in figure 2. *INT* = Internalizing; *EXT* = Externalizing; *INTp / EXTp* = Parent specific factors; *INTs / EXTs* = Self-report specific factors; *INTt / EXTt* = Teacher specific factors; *CBCL* = Child Behavior Checklist; *YSR* = Youth Self Report; *TRF* = Teacher Report Form.

Recently, the CT-C(M-1) model has been presented as a better alternative in the situation of structurally different methods. Structurally different refers to the fact that the different informants cannot be regarded as interchangeable methods, but represent structurally different points of views. Eid et al. (2008) argue that in this situation it is necessary to start by choosing a reference method and subsequently estimate other method factors as discrepancies from this reference. This approach is illustrated in figure 4, in which self report is used as reference method. For the other informants, parent and teacher in this case, correlated INT and EXT method factors are estimated. The overall INT and EXT factors will represent the covariance based on self report and the covariance in parent and teacher reports that can be predicted on the basis of self report. These factors are uncorrelated to the method factors, and therefore the method factors represent discrepancies from the reference. The

apparent drawback of this method is that it is necessary to choose a reference method, which is somewhat arbitrary if there is no gold-standard measure. We chose to fit three CT-C(M-1) models, so each informant was the reference method in one of these models.

The CT-C(M-1) models result in multiple estimations of IE-correlation, none of which can be assumed to be fully free of observation and response bias. Together with the CT-CU model, which overcorrects for bias, they will yield a range of estimated IE-correlations.

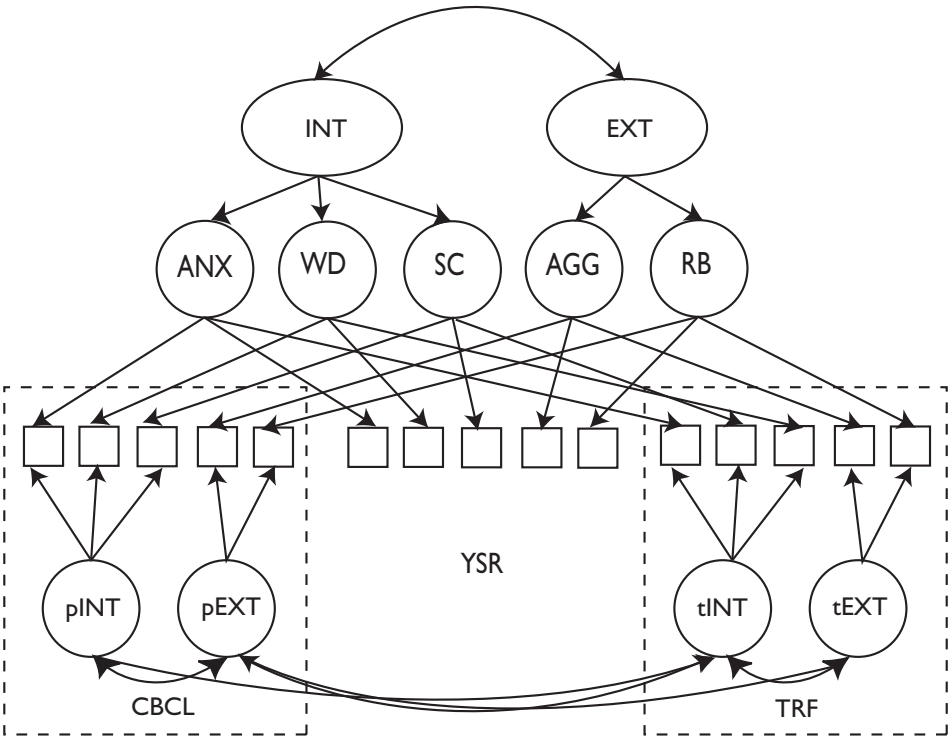

Figure 4. Correlated traits – Correlated Methods - 1 model with three informants. In this case the self-report is used as the reference method.
Note: INT = Internalizing; *EXT* = Externalizing; *INTp / EXTp* = Parent specific factors; *INTt / EXTt* = Teacher specific factors; *ANX* = Anxious-Depressed; *SC* = Somatic Complaints; *WD* = Withdrawn-Depressed; *AGG* = Aggressive Behavior; *RB* = Rule-Breaking Behavior; *CBCL* = Child Behavior Checklist; *YSR* = Youth Self Report; *TRF* =Teacher Report Form.

Instrument discrepancies

To distinguish between instrument specific and cross instrument IE-correlation, we fitted a model that is comparable to the CT-C(M-1) method described above. Given that only one of the instruments, the YSR, measures both INT and EXT it was not possible to estimate a CT-CU model. This CT-C(M-1) model is illustrated in figure 5. The covariance between YSR subscales is modeled as a specific method factor. Because of this the ASBQ and RCADS, for which no method factors can be estimated, are chosen as 'reference method'. The effect of this way of modeling is that the covariance that is specific to YSR subscales does not influence the estimated overall IE-correlation. Of course, the estimated overall IE-correlation is dependent on the ASBQ and RCADS and cannot be regarded as bias free. However, the CT-C(M-1) model offers a helpful method to compare the results of multiple versus single instrument estimations of IE-correlation, just as it does for multiple informants. If a large amount of IE-correlation is specific to the YSR, the overall IE-correlation will be small.

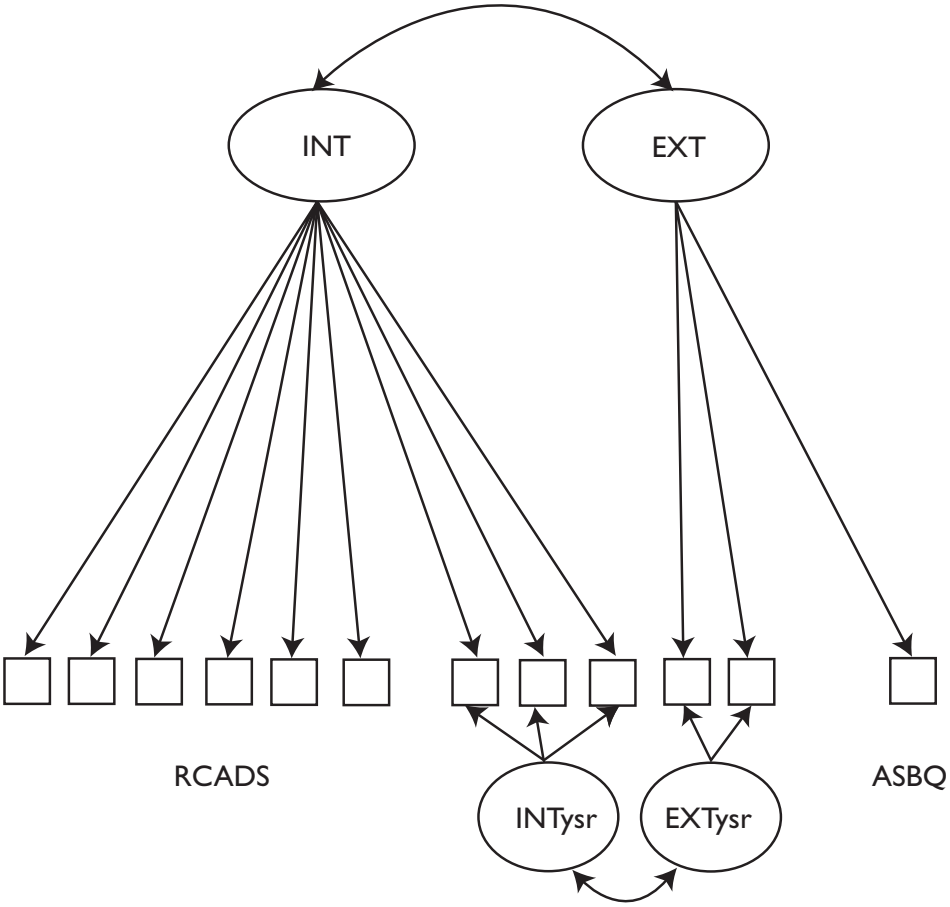

Figure 5. Correlated traits – Correlated Methods - 1 model with multiple instruments. Covariance specific to the Youth Self Report subscales is modelled as a specific method-factor.

Note: The open squares at the bottom of the figure refer to the subscales of the instruments. *INT* = Internalizing; *EXT* = Externalizing; *INTysr / EXTysr* = YSR specific.

factors; *RCADS* = Revised Child Anxiety and Depression Scale; *ASBQ* = Anti-Social Behavior Questionnaire; *YSR* = Youth Self Report.

Results

For all analyses we fitted models of Internalizing and Externalizing problems as illustrated in figures 1-5. Fit indices of the original models are reported in table 1. Also reported in table 1 are fit indices for exploratively improved models. The resulting IE-correlations and their 95% confidence intervals are reported in table 2.

Table 1. *Fit-indices of the multiple models that were fitted in order to study discrepancies.*

Discrepancy	Model	Original		Revised	
		RMSEA	CFI	RMSEA	CFI
Attrition bias	Reference	.037	.997		
	Multiple Imputation	.118	.903	.059	.985
Berksonian bias	Listwise Deletion	.131	.906	.052	.991
	Clinical Cohort	.028	.995		
	Subgroup Mental Care	.029	.994		
Stratification by age and gender	Aggregate	.062	.978		
	Stratified	.074	.968	.055	.983
Informant discrepancies	CT-CU	.080	.909	.056	.956
	CT-C(M-1)-Self	.083	.912	.059	.957
	CT-C(M-1)-Parent	.078	.921	.054	.963
	CT-C(M-1)-Teacher	.070	.936	.046	.974
Instrument Discrepancies	CT-C(M-1)-YSR	nc		.056	.966

Note: *RMSEA* = Root Mean Standard Error of Approximation; *CFI* =Comparative Fit Index; CT-*CU* = Correlated Traits – Correlated Uniqueness; *CT-C(M-1)-x* = Correlated Traits – Correlated Methods with method x as reference method; *YSR* = Youth Self Report.

Table 2. *Estimation of correlation between Internalizing and Externalizing problems for the multiple models of discrepancies.*

Discrepancy	Model	rIEOriginal[a]	rIE Revised	95% CI
	Reference	.62		.58 - .66
Attrition bias	Multiple Imputation	.64	.66	[b]
	Listwise Deletion	.65	.65	.60 - .69
Berksonian bias	Clinical Cohort	.60		.53 - .70
	Subgroup Mental Care	.67		.50 - .82
Stratification	Aggregate	.62		.58 - .65
by age and gender	Stratified	.56-.66	.56-.66	.47 - .75
Informant	CT-CU	.16	.16	.00 - .31
discrepancies	CT-C(M-1)-Self	.59	.61	.51 - .70
	CT-C(M-1)-Parent	.48	.49	.42 - .57
	CT-C(M-1)-Teacher	.37	.39	.29 - .48
Instrument	CT-C(M-1)-YSR	[c]	.31	.26 - .36
Discrepancies				

Note: *rIE* = Correlation between Internalizing and Externalizing Factors; *CT-CU* = Correlated Traits – Correlated Uniqueness; *CT-C(M-1)-x* = Correlated Traits – Correlated Methods with method x as reference method; *YSR* = Youth Self Report.

[a] 'Original' refers to the models that were developed before evaluating them with Confirmatory Factor Analysis. 'Revised' refers to models that were changed in order to improve model fit.

[b] The Multiple Imputation procedure of MPlus does not allow to compute confidence intervals.

[c] The original model was not converging.

Table 3. *Factor-loadings for the reference model: self-report at T1.*

		Reference model		Multiple imputation [a]	Listwise Deletion		Stratified [b]	
EXT BY	Anx	1.00	*0.84*	1.00	1.00	*0.71*		
	Sc	0.78	*0.58*	0.84	0.85	*0.83*		
	Wd	0.98	*0.77*	0.55	0.53	*0.56*		
INT BY	Agg	1.00	*0.97*	1.00	1.00	*0.94*	1.00	*1.00*
	Rb	0.52	*0.70*	0.53	0.50	*0.75*	0.73	*0.74*

Note: Standardized coefficients are shown in italics; *Anx* = Anxious-Depressed; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed; Agg = Aggressive Behavior; *Rb* = Rule-Breaking Behavior; *INT* = Internalizing; *EXT* = Externalizing.

[a] MPlus multiple imputation procedure does not allow to compute standardized loadings.

[b] Loadings on EXT were estimated separately for the T3 male and female subgroups.

Reference model

The reference model was based on self reported problems in the general population at T1 and is shown in figure 2. This model showed adequate fit to the data (RMSEA=.037, CFI=.997) and a substantial IE-correlation (r=.62). Factor loadings of this model are reported in table 3.

Sample discrepancies

To investigate attrition bias, the parent report based model was fitted to T3-data that were imputed using NORM. This model was compared to listwise deletion. Both procedures did not result in well fitting models (see table 1). Model fit significantly improved by freely estimating the factor loadings for both the 'listwise deletion model' (RMSEA=.052, CFI=.991) and the 'multiple imputation model' (RMSEA=.059, CFI=.985). The factor-loadings of this revised model are reported in table 3. The models hardly differed regarding IE-correlation (r=.66 versus .65), implying that attrition did not result in under- or overestimation of IE-correlation.

To investigate Berksonian bias, the self report based model was fitted on data from the clinical cohort and on the subgroup of children who received mental health care last year. Adequate fit indices were found for both the 'Clinical Cohort model' (RMSEA=.028; CFI=.995) and the 'Subgroup Mental Care model' (RMSEA=.029; CFI=.994). IE-correlations did not differ much between these samples (r=.60 and .67) and the reference model (r=.62). This indicates that Berksonian bias did not result in overestimation of self reported IE-correlation.

To investigate stratification by age and gender, the reference model was fitted in the aggregated sample (T1-T3) and in 6 subgroups (3 age-groups by 2 gender-groups). For the aggregate sample the model fit was quite adequate (RMSEA=.062, CFI=.978), but for the stratified sample fit indices were inadequate (RMSEA=.074, CFI=.968). Freely estimating the factor loadings on the EXT-factor in both the male and female T3 subgroups resulted in a better model fit (RMSEA=.055, CFI=.983; loadings reported in table 3). As shown in table 2, IE-correlations did not differ substantially between these models and were very comparable to the estimated IE-correlation in the reference model (r=.62).

Altogether, estimation of IE-correlation in different samples and subgroups did not reveal important discrepancies and all estimations were close to the .62 found for the reference model.

Informant discrepancies

To investigate informant discrepancies we fitted models including all three informants. The CT-CU model (figure 3) did not fit well to the data (RMSEA=.084, CFI=.913). The model was improved on the basis of Modification Indices provided by MPlus, which indicated some specific covariance between pairs of informants (self-parent, parent-teacher) with regard to the Wd and Sc subscales. Adding correlated

residuals between these informant specific Wd and Sc scales resulted in adequate fit indices (RMSEA=.056, CFI=.956). The factor-loadings and correlations of this model are reported in table 4. The estimated multi informant IE-correlation in this model was low (r=.16) and the 95% confidence interval even included the value of zero.

Table 4. *Factor-loadings, correlated residuals and correlations for the revised CT-CU model.*

		Self		Parent		Teacher	
EXTis [a]	anx	1.00	*0.88*	1.00	*0.85*	1.00	*0.87*
	sc	0.76	*0.62*	0.83	*0.67*	0.41	*0.50*
	wd	0.87	*0.68*	0.47	*0.51*	0.91	*0.60*
INTis	agg	1.00	*0.90*	1.00	*1.00*	1.00	*1.00*
	rb	0.50	*0.73*	0.33	*0.77*	0.43	*0.79*
		INT		EXT			
	parent [b]	1.00	*0.71*	1.00	*0.55*		
	self	0.54	*0.47*	0.63	*0.54*		
	teacher	0.64	*0.53*	1.28	*0.69*		
				Correlations			
Correlated Residuals	Parent Sc with Teacher Wd			0.21			
	Parent Wd with Teacher Sc			0.23			
	Child Sc with Parent Wd			0.20			
Correlated Methods	Self			0.63			
	Parent			0.41			
	Teacher			0.30			

Note: Standardized loadings are shown in italics. *Anx* = Anxious-Depressed; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *Rb* = Rule-Breaking Behavior; *INT* = Internalizing; *EXT* = Externalizing.

[a] 'EXTis' and 'INTis' refer to the informant specific (is) EXT and INT-factors, which are estimated on the basis of the subscales of the same informant.

[b] Loadings refer to the loadings of the informant specific (is) INT and EXT factors on the higher-order cross-informant INT and EXT factors.

Subsequently, we fitted three CT-C(M-1) models by removing one of the method factors as is illustrated in figure 4. In figure 4 self report is used as the reference method and therefore not modelled as a method factor, as can be understood by comparing figure 3 and 4. The CT-C(M-1) models did not fit well to the data (see table 1). They could be improved by adding the same correlated residuals as was done the CT-CU model. This resulted in quite adequate fit indices for the CT-C(M-1) model with self report (RMSEA=.059, CFI=.957), parent report (RMSEA=.054, CFI=.963) as well as teacher report (RMSEA=.046, CFI=.974) as reference method.

Table 5. *Factor-loadings, correlated residuals and factor correlations for the revised Correlated Traits- Correlated Methods - 1 model with Self-report as reference method.*

		Parent		Teacher	
EXTis [a]	anx	1.00	*0.79*	1.00	*0.83*
	sc	0.86	*0.65*	0.43	*0.49*
	wd	0.47	*0.48*	0.94	*0.60*
INTis	agg	1.00	*0.95*	1.00	*0.94*
	rb	0.33	*0.74*	0.43	*0.74*

		ANX [c]	SC	WD	AGG	RB
'Traits' [b]	Self	1.00 *1.00*	1.00 *0.63*	1.00 *0.70*	1.00 *1.00*	1.00 *0.81*
	Parent	0.33 *0.31*	0.52 *0.26*	0.21 *0.17*	0.50 *0.34*	0.36 *0.30*
	Teacher	0.44 *0.34*	0.19 *0.17*	0.26 *0.14*	0.43 *0.30*	0.55 *0.36*

		Higher-order [d]	
INT	ANX	1.00	*0.86*
	SC	0.78	*1.00*
	WD	0.91	*1.00*
EXT	AGG	1.00	*0.99*
	RB	0.41	*0.81*

		Correlations
Correlated residuals	Parent Sc with Teacher Wd	*0.20*
	Parent Wd with Teacher Sc	*0.23*
	Child Sc with Parent Wd	*0.20*
Correlated Methods	INTp with EXTp [e]	*0.57*
	INTt with EXTt	*0.44*
	INTp with INTt	*0.34*
	EXTp with EXTt	*0.36*
	INTp with EXTt	*0.16*
	EXTp with INTt	*0.13*

Note: Standardized loadings are shown in italics; *Anx* = Anxious-Depressed; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *Rb* = Rule-Breaking Behavior; *INT* = Internalizing; *EXT* = Externalizing; *INTp / EXTp* = parent-specific factors; *INTt / EXTt* = teacher-specific factors.

[a] 'EXTis' and 'INTis' refer to the informant specific (is) EXT and INT-factors, which are estimated on the basis of the subscales of the same informant.

[b] Estimations of 'Traits' are based on reports of the three informants on the same subscale. The term is placed between apostrophes because in the CT-C(M-1) model they are dominated by the reference method.
[c] The acronyms in capitals refer to factors with loadings of the same subscale as reported by all informants. So, for example ANX refers to a factor on which the Anx subscales reported by child, parent and teacher load.
[d] Shown are the loadings of the five 'traits' on the higher order INT and EXT factors.
[e] In the CT-C(M-1) model the method factors are allowed to be correlated. So, for example, the parent-specific INT factor (INTp) can be correlated to both the parent specific EXT factor (EXTp) and the teacher-specific INT factor (INTt) and EXT factor (EXTt).

The factor-loadings and correlations for the CT-C(M-1) model with self-report as reference method are shown in table 5. As can be observed in table 5, substantial correlations were found between the informant-specific INT and EXT factors and low loadings were found of the non-reference methods (parent and teacher reports) on the trait-factors. These observations converge with the findings in the CT-CU model that most IE-correlation can be attributed to unique informant reports. The estimated IE-correlation differed substantially depending on the chosen reference method and was highest for self report (r=.61), lower for parent report (r=.49) and lowest for teacher report (r=.39).

Instrument discrepancies

To investigate instrument discrepancy, the CT-C(M-1) model illustrated in figure 5 was fitted to the data. In this model we assume YSR specific IE-correlation on the one hand and multi instrument correlation on the other. The model showed in figure 5 did not result in a convergent model, which appeared to be caused by a misconstruction of the YSR specific EXT-factor. This problem could be solved by separately estimating the correlation between YSR-INT and the subscales Rb and Agg, rather then estimating a YSR-EXT factor. Furthermore, to find a satisfactory model-fit (RMSEA=.056, CFI=.966) it was necessary to add a residual correlation between the RCADS Depression scale and the YSR Wd scale. Factor-loadings and correlations of this revised model are shown in table 6.

It was found that some IE-correlation may be specific to the YSR-instrument and specifically to the correlation between the YSR Agg scale and Internalizing problems. The estimation of multi instrument IE-correlation was lower than in the reference model (r=.31 versus r=.62). This indicates that the estimated IE-correlation is dependent on the instruments that are used.

Table 6. *Factor-loadings, correlated residuals and factor correlations for the revised Correlated Traits- Correlated Methods - 1 model for instrument discrepancy.*

Factor	Subscale	Factor loadings	
EXT	Rb	0.15	*0.88*
	Agg	0.18	*0.76*
	ASBQ	0.25	*0.73*
INT	RCanx	0.34	*0.74*
	RCdep	0.24	*0.74*
	RCocd	0.33	*0.74*
	RCpd	0.27	*0.75*
	RCsp	0.32	*0.75*
	RCsa	0.24	*0.69*
	Anx	0.20	*0.73*
	Sc	0.16	*0.51*
	Wd	0.16	*0.56*
INTysr	Anx	1.00	*0.47*
	Sc	0.66	*0.27*
	Wd	1.14	*0.50*

			Correlations
INT	with	EXT	*0.31*
INTysr	with	Rb	*0.59*
INTysr	with	Agg	*0.59*
INT	with	Agg	*0.29*
Wd	with	Rcdep	*0.24*

Note: Standardized loadings are shown in italics. *Anx* = Anxious-Depressed; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *Rb* = Rule-Breaking Behavior; *ASBQ* = Anti-Social Behavior Questionnaire; *RC…* = Subscale of the Revised Child Anxiety and Depression Scale (RCADS); *RCanx* = Generalized Anxiety Disorder; *RCdep* = Major Depressive Disorder, *RCocd* = Obsessive Compulsive Disorder; *RCpd* = Panic Disorder; *RCsp* = Social Phobia; *RCsa* = Separation Anxiety Disorder; *INT* = Internalizing; *EXT* = Externalizing; *INTysr* = Internalizing factor uniquely measured by YSR-subscales.

Discussion

In the current study the hypothesis that the correlation between Internalizing and Externalizing problems results from methodological artefacts could not be completely rejected, because in one model (the CT-CU model) all IE-correlation could be attributed to informant specific factors. This model probably overcorrects bias and therefore the results should not be interpreted as strong evidence that comorbidity between the Internalizing and Externalizing domain is nothing but an artefact. However, the results do support the idea that using only a single informant and a single instrument can easily result in overestimation of IE-correlation. Furthermore,

the multi instrument and multi informant models reveal that estimated correlations between the Internalizing and Externalizing domain in part result from several instrument- and informant-specific sources.

In the TRAILS longitudinal multi cohort study of (pre)adolescents, sampling bias did not seem to result in excess correlation between Internalizing and Externalizing problems. Attrition bias and population stratification by age and gender did not or only slightly influence the IE-correlations, and Berksonian bias did not result in an overestimation: the IE-correlation was similar in clinical subgroups and the general population sample.

Informant discrepancy was found to be large. This finding is in line with Youngstrom et al. (2003), who showed a very low prevalence of comorbid Internalizing and Externalizing disorders when the AND-rule (i.e. all informants must rate a symptom as present) was applied to data of parent, child and teacher, but a large prevalence when the OR-rule (i.e. only one of the informants must rate the symptom) was applied. These informant discrepancies are probably related to informant specific processes of observing and responding, which may include observation and response biases. However, informant specific processes cannot be equated to bias. The alternative provided by the CT-C(M-1) models resulted in much higher estimations of IE-correlation. However, these estimations are dependent on the chosen reference method, i.e. on the perspective of one informant. The resulting IE-correlations are almost equal to what is found when only one informant is consulted and may still contain observation and response biases. The conclusion that can be drawn from both the CT-CU and the CT-C(M-1) models is that the estimation of IE-correlation is strongly informant dependent.

The strong informant dependency implies that explanations for comorbidity may not only involve hypotheses on the causal structure of symptoms and disorders, but also on the process of observing a subject and responding to questionnaires. Moreover, to fully understand the constructs that are being measured and to fully assess their validity it would be necessary to give an account of the process that causes responses on questionnaires. In our view, the CT-C(M-1) model provides a useful starting point for developing models that incorporate informant specific perspectives and discrepancies.

Which informant is selected as reference method is very important as different informants yield different results. Specifically, when the teacher was used as the reference method in the CT-C(M-1) model, the estimated IE-correlation was quite low. This discrepancy may be related to a teacher specific process of observing and responding, for example a tendency to contrast and 'categorize' children in a classroom or a stronger sensitivity to externalizing behaviors. An alternative hypothesis to explain the findings is that the discrepancy does not result from observation and response processes, but from an actual difference in behavior of the (pre)adolescent in different contexts. If a (pre)adolescent displays some problems in one context and other problems in another context, i.e. if problems are context

dependent, we may expect a lower correlation between problems if reports are given by an informant who mainly observes the (pre)adolescent in one context. Self report can be based on a broad sample of behaviors in different contexts, while the teacher mainly observes in the classroom or at the schoolyard.

We also found discrepancies between instruments, which suggests that some IE-correlation was specifically related to the Youth Self Report. This finding does not necessarily imply that IE-correlation is overestimated by the YSR, because it may also be underestimated by the other instruments. In general these results show that IE-correlation is dependent on the way Internalizing and Externalizing problems are defined and measured. More specifically, some Internalizing and Externalizing problems may co-occur often, while others may be relatively independent. The Aggressive Behavior Scale of the YSR showed a particularly strong correlation with Internalizing problems. A face value evaluation of the content of this scale shows that some items (e.g. mood-swings) are conceptually related to both the Internalizing and Externalizing domain. Such conceptual analysis is beyond the scope of the current paper, but may proof crucial to better understand the instrument discrepancies that we found.

This paper illustrates that analysis of discrepancies can contribute significantly to an understanding of comorbidity and measurement of psychopathology. The results point to a number of interesting directions for future research. First, research may be devoted to a better understanding of the causes and consequences of informant discrepancy. One possibility is including reports of more informants (peer ratings, siblings, second parent) or information about informant characteristics (e.g. parental depression, see De Los Reyes, Goodman, Kliewer, & Reid-Quinones, 2008). Qualitative interviews may be useful to understand how informants observe a subject and complete questionnaires. Second, we regard the CT-C(M-1) models employed in this paper as useful tools in developing our understanding of informant discrepancy and therefore agree with the recommendation by Eid et al. (2008) to use these models in those cases where multiple structurally different methods (like parent, child and teacher) are used. Structurally different means that the methods are not interchangeable, but rather provide a unique perspective. Third, future research designs may tackle the issue of discrepant findings for different instruments. One way to proceed is to develop and test the idea that some Internalizing and Externalizing problems co-occur above chance expectation, while other items may almost only be correlated due to bias. This kind of analysis, although not providing direct evidence of causality, may eventually result in a better understanding of specific relations between Internalizing and Externalizing problems. Such specificity is needed to uncover the latent mechanisms that cause comorbidity between disorders.

Chapter 6

Stability and predictive utility of differences between informants

Abstract

Differences between informants challenge the quality of measurement and diagnosis, but might also provide useful diagnostic information. The current study was aimed at testing whether differences between informants showed rank-order stability during adolescence. Furthermore, it was tested whether these differences predicted changes in self-reported problems over time. The results showed large rank-order stability of the deviance of parent-reported problems from self-reported problems. Differences between informants were predictive of change in self-reported problems over time. These associations were probably too weak to be very useful for predictive purposes in clinical practice. It is discussed how the very stable deviances between self- and parent-reported problems that were found may be used in clinical practice for other purposes than prediction.

Introduction

A common finding in research on child- and adolescent psychopathology is that problem scores derived from different informants are only weakly correlated. This indicates that one or more of the instruments is unreliable, or that the instruments do not validly measure the same construct (Courvoisier, Nussbeck, Eid, & Cole, 2008) or a combination of these. If instruments are unreliable then each measurement can be regarded as an imprecise piece of information about the same construct. If they are not valid measures of the same construct, then each measurement may be regarded as a piece of information about a different construct. Related to this second possibility, it has been argued that information from each informant can be regarded as adding specific diagnostic information (Achenbach, et al., 1987). Even discrepancies between different informants as such may contribute to a more complete picture of an individual's functioning. To test the relevance of this perspective several questions need to be addressed, two of which are the focus of the current article. The first question is how discrepancies develop over time. Do they show stability or do they vary across measurement occasions? The second question is how useful discrepancies are in terms of predictive value for the development of self-reported behavioral and emotional problems.

Specific hypotheses with regard to the above-mentioned questions can only be developed if we have an idea of what is actually measured by discrepancies. The most general and obvious feature of informant discrepancies is that they demonstrate that

multiple answers can be given to the same questions about the same individual. We distinguish four specific aspects that are likely to influence this variability. First, differences between informants can indicate uncertainty about the answer to a question, so that informants may make different guesses. Second, differences can indicate that the actual behavior to which the questions refer is variable and only expressed in certain contexts or at certain moments (e.g. Kraemer, et al., 2003). Third, differences can indicate that some behaviors, for instance emotions that are not clearly expressed, are difficult to observe for some informants (e.g. Harkness, Tellegen, & Waller, 1995). Fourth, differences can indicate that informants have a different judgment about the same observations (see De Los Reyes & Kazdin, 2005). In sum, we assume that differences between informants indicate informant uncertainty, variability of behavior, low observability, or differences in judgment. In our view, these four general terms cover most of the more specific causal pathways that have been described in the literature. For example, differences between self- and parent-report may be related to denial of problems (Harkness, et al., 1995) or oppositionality, which can be regarded as differences in judgment or visibility. Also, it has been suggested that differences may be related to differences in the context in which a subject is observed (De Los Reyes & Kazdin, 2005), which can be regarded as involving variability of behavior over contexts. Finally, the four terms also include idiosyncratic reasons for differences between informants, which can be viewed as informant-specific interactions between observations, judgments and responses.

Variability and low observability are behavior properties that may change during adolescence. For example, observability may increase due to communication between parents and adolescents. It may also decrease due to the fact that an adolescent spends less time at home. However, at the same time we can expect stability: if a behavior-pattern is very context-dependent or difficult to observe for others at one measurement occasion, it is likely that the same behavior-pattern will also be more variable over context or difficult to observe at a later measurement occasion. Furthermore, the context and perspective of a specific informant and the manner in which informants judge about behaviors, can be expected to have some stability as well. For these reasons, informant discrepancies are expected to show a substantial amount of stability. This hypothesis is supported by previous reports on the stability of informant-specific factors and differences between informants in childhood (e.g. Courvoisier, et al., 2008; Grimm, Pianta, & Konold, 2009).

With regard to the question how useful informant discrepancies are as a predictor of the course of self-reported problems, we expect larger differences to be associated with more variability over time. If a response is given with uncertainty, the question is apparently not very clear, which increases the likelihood that the respondent will give a different response at a future occasion. If the behavior is variable, it is relatively likely to change over time due to changes in context like going to another school. This is especially true in adolescence, where such changes are common. If there are

differences in judgment, there is some chance that the self-report will change because of the influence of the judgment of others.

With respect to the direction of changes over time, we hypothesize that in case of informant discrepancies, self-reported ratings will change in the direction of other-reported ratings. More specifically, if self-reported scores are low compared to other-reports, we expect them to increase, and if they are relatively high, we expect them to decrease over time. If discrepancy indicates a range of possible responses, it is more likely that a response will change into an alternative that has been given previously by another informant than that a completely new response will be given. At the same time, we expect that more discrepancies, regardless of whether the other informant reports more or less, predict a decrease of self-reported problems over time. Thus, we expect *non-directional* discrepancies to predict a decrease of self-reported problems and we predict *directional* discrepancies to predict a change towards the amount of problems reported by the other informant. The rationale for this is that we assume that discrepancies indicate uncertainty about and variability of the problems that the items ask about. We expect that these problems are more likely to involve temporary, context-dependent, issues rather than stable traits and are therefore more likely to disappear at later ages.

The predictive value of informant discrepancies has been suggested by results of Pelton and colleagues (Pelton & Forehand, 2001; Pelton, Steele, Chance, Forehand, & Family Hlth Project Res, 2001), who reported that differences between self- and parent-reports on the quality of the parent-child interactions predicted internalizing and externalizing problems at a later age. Furthermore, differences between self- and parent-reported symptoms have been found to predict negative outcomes like problems at school, police contacts and drug abuse (Ferdinand, van der Ende, & Verhulst, 2004, 2006). Others have reported that differences between self- and parent-reported parenting style were predictive of depressive symptoms and anti-social behavior (Feinberg, Howe, Reiss, & Hetherington, 2000), and of internalizing problems and social competence (Guion, Mrug, & Windle, 2009).

In sum, we aimed to investigate several expectations with respect to the stability and predictive utility of differences between informants. These expectations were tested on data from a longitudinal study of (pre)adolescents. We used self- and parent-reports of behaviors and emotions related to aggression on the one hand and withdrawal from social contact and depression on the other. For both domains we expected to find stable differences between informants and utility of these differences in predicting variability in self-report over time.

Methods

Sample

Subjects were participants in the 'Tracking Adolescents' Individual Lives Survey' (TRAILS), a prospective multi-cohort study of Dutch (pre)adolescents. The study

involved a representative sample from the general population and is described in detail in Huisman et al. (2008). Briefly, the target sample involved all 10- to 11-year-old children living in the three largest cities and some rural areas in the North of The Netherlands. Of the eligible children, 76.0% (n=2230, mean age = 11.09, SD =0.55) were enrolled in the study. Responders and non-responders did not differ regarding the prevalence of teacher-rated problem behavior and associations between sociodemographic variables and mental health indicators (De Winter, et al., 2005). To date, the population cohort has been assessed three times (T1: March 2000- July 2001, T2: September 2003- December 2004, T3: September 2005-December 2007). Participation rates were 96.4% at T2 (mean age= 13.55, SD = 0.53), and 81.4% at T3 (mean age= 16.25, SD = 0.73). After complete description of the study to the subjects, written informed consent was obtained from the parents at each assessment wave and from the adolescents at T2 and T3. T1, T2, and T3 data are used in the present study.

Instruments

The Dutch versions of the Youth Self-Report (YSR; Achenbach, 1991c; Verhulst, et al., 1997b) and the Child Behavior Checklist (CBCL; Achenbach, 1991a; Verhulst, et al., 1996) were used to assess self- and parent-reported behavioral and emotional problems. The CBCL and YSR are 112-item questionnaires on which emotions and behaviors are rated on a 3-point rating scale (not [0], sometimes [1], or very often [2]). The period over which respondents are asked to report is the last six months. In the TRAILS-study the questionnaire was completed by one of the parents, which was the mother in most cases. Factor analysis on these items revealed a structure of eight syndrome scales (Achenbach, 1991c). In the current study the scales Withdrawn-Depressed (Wd; 8 items, CBCL: α=0.71, YSR: α=0.64) and Aggressive Behavior (Agg; 18 items, CBCL: α=0.88, YSR: α=0.82) will be used.

Statistics

All analyses were done using MPlus version 5.2. Because Wd and Agg are not normally distributed we used maximum likelihood estimation with robust standard errors (MLR), which is relative robust to deviations from normality. We used the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI) to evaluate model fit. An RMSEA below .05 and a CFI above .95 were regarded as indicating adequate fit.

• *Defining differences between informants*

We used two distinct measures of the differences between informants. On the one hand we developed a measure of the absolute amount of discrepancies that are observed between informants (DISi). The DISi scores were based on a summation of all discrepancies between responses on all items of a subscale:

$$DISi = \Sigma_{i=1\ldots s} (|\text{Item i self} - \text{Item i parent}|).$$

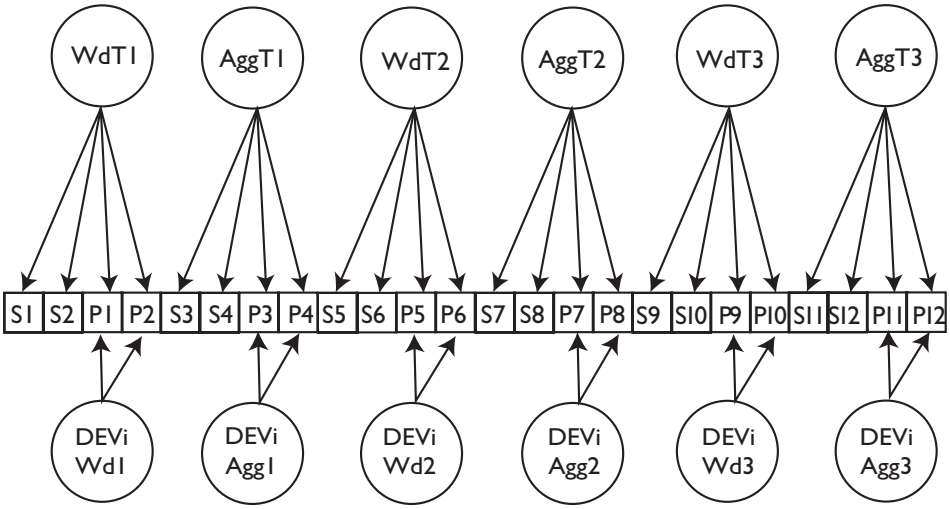

Figure 1. The CT-C(M-1) model for three occassions of measurement of subscales Wd and Agg.

Note: Correlations between factors are excluded for simplicity. All factors are correlated except for self-report and parental report at each measurement wave (T1-T3). *Wd*=Withdrawn-Depressed; *Agg* = Aggressive Behavior; *DEVi* = deviance of other informant (parent); *S1-S12* indicators of self-reported problems; *P1-P12* indicators of parent-reported problems.

On the other hand we used a measure of the 'deviance' of the report of another informant from self-reported problems (DEVi) on the basis of a recently developed latent model for multi-informant data. This measure is high when the other informant reports more problems, and low when the other reports less problems. The DEVi scores were derived from the "Correlated Traits – Correlated Methods minus one model" (CT-C(M-1); Eid, et al., 2003). Overviews of this model and arguments for why it is a good approach for modeling multi-informant data are given elsewhere (Courvoisier, et al., 2008; Eid, et al., 2008). For the current article this model (see figure 1) was used to estimate the unique variance of parental reports, i.e. the information that is not predictable from self-report. To develop a CT-C(M-1) model one has to choose a reference-method, which in this case was self-report. Subsequently, one estimates 'Trait' factors on which indicators of all informants have

loadings, and 'Method' factors on which indicators of only one informant have loadings. These 'Method' factors are estimated for all informants except for the reference-method and are uncorrelated to the 'Trait' factors. Therefore, these factors are best interpreted as unique deviance of other informants from the reference-method (Geiser, Eid, & Nussbeck, 2008). In the current article we chose self-report as a reference-method and used the CT-C(M-1) model to estimate deviance scores of parental report (DEVi). Unfortunately, fitting the model on the basis of items as indicators resulted in a too complicated and non-converging model. As an alternative we used a split half method (see Courvoisier, et al., 2008). We computed indicators on the basis of half of the items of each subscale, which resulted in two indicators for each subscales at each assessment wave. These indicators were used to fit the CT-C(M-1) model.

Our DISi measure reflects all observed discrepancies on a subscale regardless of the direction of these discrepancies. Earlier approaches to define non-directional discrepancy have used the absolute difference of scale scores rather than items (Feinberg, et al., 2000; Pelton & Forehand, 2001). This approach is likely to correlate with ours, but in our view the item-level approach is more consistent in capturing all non-directional discrepancies between informants. Low scores indicate few discrepancies, high scores indicate many discrepancies. It appears that the DISi scores are positively correlated to the variables estimated in the CT-C(M-1) model. This can be understood as indicating that if more problems are observed the amount of discrepancy will generally be higher. However, variance in DISi cannot be completely explained by the CT-C(M-1) model. As the CT-C(M-1) approach provides a concise latent variable model for multi-informant data we were only interested in the variance in DISi that was not explained by the CT-C(M-1) model. Therefore, in all analyses of DISi we controlled for the influence of the CT-C(M-1) factors.

• *Defining variability over time*

We used two different variables to capture differences in self-reported problems between multiple occasions of measurement (T1, T2 and T3). First, we used a linear growth model (see figure 2). In a linear growth model an individual growth curve is estimated for each individual, which is characterized by two variables: an intercept and a slope. We specified the slope as the linear deviance from the T1-measure, which will be referred to as deviance over time (DEVt). High values indicate an increase of self-reported problems over time, while low values indicate a decrease.

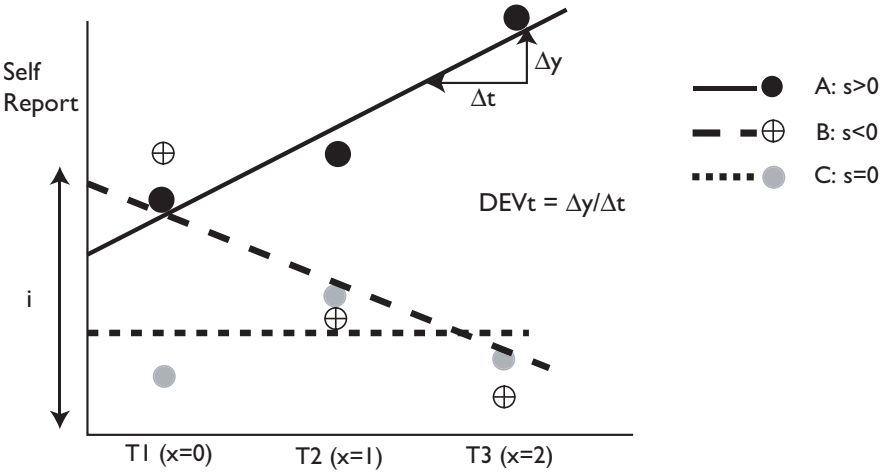

Figure 2. Examples of the individual growth-curves of three individuals (A-C).
Note: i = intercept; *DEVt* = deviance over time; *T1-T3* = measurement waves; *A* = individual with increase of self-reported problems over time; *B* = individual with decrease; *C*=individual with no linear change.

The DEVt variable can be used to investigate whether differences between informants are predictive of increases or decreases of self-reported problems over time. However, we were also interested in whether differences between informants were predictive of any variability over time, regardless of the direction of this variation. For this reason we developed a variable indicating the amount of 'discrepancy over time' in item-responses between T1 and the other occasions of measurement:

$$DISt = \Sigma_{i=1\ldots s} (|Item\ i\ T1 - Item\ i\ T2| + |Item\ i\ T1 - Item\ i\ T3|).$$

High values indicate many discrepancies between self-reported problems over time, while low values indicate few discrepancies.

• *Stability and predictive utility*

We investigated the rank-order stability of the variables DEVi and DISi over the three occasions of measurement by computing test-retest correlations between different occasions of measurement (T1, T2, T3). Subsequently, we used linear regression analysis to test whether DEVi and DISi predicted DEVt and DISt. First, we expected DISi to be a predictor of DISt. That is: we expected that more discrepancies between informants at T1 would predict more discrepancies between T1 self-reported

problems and self-reported problems at T2 and T3. Second, we expected DEVi to predict DEVt, because we expected that most change would occur in the direction of the problems reported by the other informant. That is: we expected that if parents reported more problems at T1 then self-report would increase over time. Third, we expected DISi to predict DEVt, because we expected discrepancies to be an indicator of instability, unreliability or context-dependency of the reported behaviors. That is: we predicted that more discrepancies at T1 would predict a decrease of self-reported problems over time.

• *Preventing chance findings*

Because we fitted multiple latent variable models and tested several linear regression models we felt that there was a risk of chance capitalization. For this reason we randomly split the original sample in two parts. In the first sample (N=1099) we developed the latent variable and regression models and we replicated these models in the second sample (N=1124).

Results

The CT-C(M-1) model

The CT-C(M-1) model with self-report as reference method fitted well to the data (χ^2 =631.5; *df* = 183; RMSEA<.05; CFI > .95). Factor-loadings of this model are reported in table 1. For each measurement-wave (T1, T2, T3) and each subscale this model resulted in a distinction between reference factors, which are strongly influenced by self-report, and unique deviance factors (DEVi) which are only influenced by parental report and uncorrelated to the reference factors. The DEVi factors were calculated for both the Aggressive Behavior and the Withdrawn-Depressed subscales and were used in subsequent analyses.

Growth model of self-reported problems

A growth model was fitted that included T1-T3 self-reports of the subscales Withdrawn-Depressed and Aggressive Behavior. Factor-loadings are reported in table 2. This model showed adequate fit-indices (χ^2 =16.8; *df* = 6; RMSEA<.05; CFI>.95). The model resulted in estimated linear growth curves for each individual for both the Aggressive Behavior (Agg) and Withdrawn-Depressed (Wd) scales. These are indicated by individual-specific intercept-variables and slope-variables. The slope-variables capture individual differences in linear change during adolescence (DEVt).

Table 1. *Standardized loadings of the CT-C(M-1) model for Withdrawn-Depressed and Aggressive Behavior:*

Informant	indicator	T1 Wd	T2 Wd	T3 Wd	DEVi T1 Wd	DEVi T2 Wd	DEVi T3 Wd
Self	Wd1	0.77	0.73	0.81			
	Wd2	0.67	0.72	0.79			
Parent	Wd1	0.21	0.33	0.31	0.75	0.76	0.78
	Wd2	0.24	0.35	0.35	0.67	0.68	0.72
		T1 Agg	T2 Agg	T3 Agg	DEVi T1 Agg	DEVi T2 Agg	DEVi T3 Agg
Self	Agg1	0.86	0.84	0.84			
	Agg2	0.84	0.81	0.88			
Parent	Agg1	0.35	0.38	0.37	0.84	0.82	0.81
	Agg2	0.34	0.40	0.40	0.84	0.83	0.84

Note: The indicators are derived from splitting the subscales in halves for each informant at each assessment wave. *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *DEVi* = Deviance of parental report from self-report.

Table 2. *Standardized loadings of the growth-model for self-reported Withdrawn-Depressed and Aggressive Behavior.*

	Intercept	DEVt Wd
Wd T1	0.687	
Wd T2	0.682	0.293
Wd T3	0.637	0.548
		DEVt Agg
Agg T1	0.674	
Agg T2	0.699	0.269
Agg T3	0.678	0.523

Note: The indicators are derived from splitting the subscales in halves for each informant at each assessment wave. *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *DEVt* = linear change relative to T1.

Rank-order stability of differences between informants

The rank-order stability of differences between informants was investigated by computing test-retest correlations of the DEVi and DISi variables for both the Aggressive Behavior and Withdrawn-Depressed subscales. As reported in table 3, all test-retest correlations for the DEVi variables were above .75. This indicates that the deviance of parental report from self-report tends to be very stable during

adolescence. DISi on the other hand was much less stable, with test-retest correlations between .09 and .25. These are partial correlations, because at each measurement-wave (T1, T2, T3) we controlled for self-reported problems and DEVi. Thus, the deviance (i.e. higher or lower estimation) of parental reports from self-report was found to be very stable over time, but the amount of discrepancy in item responses much less so.

Table 3. *Test-retest correlations of differences between informants.*

Subscale	Informant difference	T1 - T2	T2 – T3	T1-T3
Wd	DISi	.15	.19	.09
	DEVi	.82	.87	.80
Agg	DISi	.25	.24	.19
	DEVi	.80	.79	.77

Note: *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *DISi* = Discrepancies between informants; *DEVi* = Deviance of parental report from self-report.

Table 4. *Differences between informants as predictors of variability in self-reported problems over time.*

Subscale	Variability	T1 predictors	B	SE	*Beta*	R^2
Wd	DISt	Self-report	5.85	0.41	*0.53**	.43[a]
		DISi	0.32	0.07	*0.20**	
		DEVi	-0.02	0.12	*0.00*	
						.46
	DEVt	Self-report	-0.17	0.06	*-0.14**	.05
		DISi	-0.02	0.01	*-0.14**	
		DEVi	0.09	0.02	*0.20**	
						.09
Agg	DISt	Self-report	12.66	0.81	*0.56**	.48
		DISi	0.41	0.06	*0.26**	
		DEVi	-0.23	0.10	*-0.07*	
						.51
	DEVt	Self-report	-0.75	0.12	*-0.24**	.09
		DISi	0.26	0.04	*-0.09**	
		DEVi	0.10	0.02	*0.22**	
						.12

Note: *Wd* = Withdrawn-Depressed; *Agg* = Aggressive Behavior; *DISt* = discrepancies in self-report between T1 and other occasions of measurement (T2, T3), *DEVt* = deviance (linear increase/decrease) of self-reported problems over time, *DISi* = discrepancies between informants; *DEVi* = Deviance of parental report from self-report.

[a] The R^2 after self-report shows the amount of explained variance if only self-report was used as a predictor. It does not correspond to the regression-coefficients, which are all derived from the analyses with multiple predictors.

* $p<.05$

Predictive utility of differences between informants

Linear regression analysis was used to test whether differences between informants (DEVi and DISi) were predictive of variability over time (DEVt and DISt). The results are shown in table 4 and were very similar for the Agg and Wd subscales. First, as expected DISi was a positive predictor of DISt. This indicates that more discrepancies in item-responses between informants at T1 predict more discrepancies of self-reported problems between T1 and later occasions of measurement (T2 and T3). Second, DEVi was a positive predictor of DEVt. This indicates that the directional deviance of parental reports from self-report was predictive of the direction of change in self-reported problems over time. Thus, as expected, higher parental reports predict an increase of self-reported problems and lower parental reports predict a decrease. Finally, DISi was a negative predictor of DEVt. This indicates that, confirming our hypothesis, more discrepancies in item responses are predictive of a decrease in self-reported problems over time.

Thus, both DEVi and DISi predicted variability in self-reported problems over time. However, the amount of explained variance was small. As can be observed from table 2, differences between informants (DEVi and DISi) did not substantially increase the amount of explained variance in variability over time (DEVt and DISt).

Replication

The above presented results were all based on analyses of a random half of the total sample. All analyses were replicated with the other half. The results replicated very well. However, one discrepancy was found (all other results can be obtained from the authors upon request). The growth-model of self-reported problems showed slightly worse model-fit in the replication sample (χ^2 =16.8; *df* = 6; RMSEA=.061; CFI>0.95). This shows that the linear growth-model did not completely capture the development of problems during adolescence, but does not invalidate the results presented in this paper because the interpretation of the slopes (DEVt) is not altered by it.

Discussion

It has been suggested that discrepancies between informants may contain useful diagnostic information (e.g. De Los Reyes, & Kazdin, 2005). The aim of the current study was to investigate the stability and predictive utility of this information. Strong support was found for rank-order stability of the deviance between parental report and self-report. Furthermore, informant discrepancies predicted variability in self-reported problems during adolescence. However, predictive utility was small: differences between informants accounted for little explained variance in variability of self-reported problems during adolescence. In the following these results will be used to evaluate whether informant differences are interesting psychological phenomena that contain useful information, and whether this information can be used in clinical practice. Before discussing these three issues we will mention some strengths and limitations of the study.

Strengths and limitations

A particular strength is that differences between informants were captured by two measures. In our view these two measures are well-suited to cover the full domain of informant discrepancies. On the one hand the CT-C(M-1) model adequately captures systematic deviances of other informants from self-report. On the other hand the summarized discrepancies between item-responses capture non-systematic non-directional discrepancies. Furthermore, we used two different measures of variability over time. We think that these two measure together adequately capture both the systematic change and non-systematic variability of self-reported problems during adolescence. Altogether, we think that these four variables allow for a rather complete understanding of the associations between informant differences and variability over time.

The generalizability of the study is obviously limited by the fact that it only covers self-reports and parental reports during adolescence in one specific longitudinal study (TRAILS) using one specific measurement approach (CBCL and YSR). Furthermore, we only evaluated the utility in predicting variability of self-reported problems over time. This outcome was useful in evaluating the diagnostic value of differences between informants, but a complete evaluation of predictive utility should capture other outcome variables as well.

Can differences between informants be regarded interesting psychological phenomena?

This study clearly showed stability of the deviance between self- and parent-reported problems during adolescence. Obviously, these impressively stable deviances deserve scientific attention. They can be completely interpreted as involving stability of the reports of each single informant. That is, each informant-report is by far the best predictor of the report of that same informant at a later occasion. This does certainly not imply that these reports are only influenced by informant-characteristics or 'bias' (van der Valk, van den Oord, Verhulst, & Boomsma, 2001).

Furthermore, it was found that differences between informants were predictive of variability in self-reported problems over time. Therefore, differences between informants at T1 may be interpreted as contributing diagnostic information about self-reported problems at T1. More discrepancies predict that the T1 self-report is somewhat more likely to change. Agreement between multiple informants was a predictor of stability, disagreement was a predictor of change.

What kind of diagnostic information is captured by differences between informants?

The study suggests that differences between informants primarily reflect the stability of each unique informant's report. This uniqueness is probably related to the specific perspective and context of that informant (De Los Reyes, 2009; Kraemer, et al., 2003; Noordhof, et al., 2008). However, the selves and parents involved are also

unique individuals with a unique way of accomplishing the task of completing a questionnaire. There are many ways to do this and the informants do not get feedback on how well they performed. All non-random idiosyncratic ways of observing and responding may contribute to a stable deviance between two unique informants. Therefore, the most important diagnostic information that is contained in differences between informants is that the instruments do not reliably measure exactly the same underlying quantitative construct. Assuming that none of the informant-reports can be accepted as a gold standard measure this implies that the instruments do not permit a very precise 'unbiased' estimation of the actual amount of psychological problems. This does not mean that the two instruments (self-report and parent-report) may not give very precise estimations of two different quantitative constructs related to two different perspectives (self vs parent).

Differences between informants were somewhat predictive of systematic change and non-systematic variability of self-reported problems over time. The associations found were in the direction expected on the basis of our hypotheses. Thus, the results do not contradict the very broad interpretations of differences between informants given in the introduction (uncertainty, variability, observability, judgment). However, the small associations found and the large amount of possible explanations for these findings do not permit strong conclusions regarding what is measured by differences between informants.

Can differences between informants be used in clinical practice?

We did not find evidence that discrepancies between informants have much utility for predicting the variability in self-reported problems during adolescence, because the amount of explained variance was too small. As discussed by McGrath (2008) in situations of practical decision-making a 'best single predictor' heuristic may often be preferable to taking into account multiple predictors of an outcome, which is specifically the case with weak predictors that add only little explained variance.

However, we do not think that the utility of differences between informants should only be evaluated in terms of predictive utility. Specifically, we think that differences may have utility both in diagnosis and in therapeutic interventions. The multiple instruments used in clinical practice do not result in a precise estimation of the actual amount of problems, but in multiple rather stable pieces of information, which may be imagined as points on a hypothetical range. This range, which will show much stability over time, may currently be a more realistic conceptualization of psychopathology than the quest for point-estimates. Accepting this viewpoint in clinical practice means that an estimate given by one informant is regarded as only one point on a hypothetical range. Adding more instruments, repeated measures and more informants does not help to reduce this range. On the contrary, it helps to further characterize the range of estimates that can be given for the same person. Communicating about ranges and uncertainties can increase the reliability of diagnostic information and using differences between informants can be one way to accomplish this goal. A very interesting and

comparable idea has been proposed for conceptualizing the range of possible change (RPC) supported by clinical trials (De Los Reyes & Kazdin, 2006).

Differences between informants may also have therapeutic utility. Using self-report as a reference-measure a clinician may discuss deviances of other reports from self-reported problems and aim to understand the idiosyncratic reasons for these differences. While some of these may just involve differences in understanding items or response-styles, other deviances may result from a real difference between contexts (e.g. home vs school), between self-judgment and judgment by others or a lack of insight by some informants. Such post-hoc individualized explanations will currently not result in explicit and quantifiable knowledge (see McGrath, 2008), but may increase diagnostic understanding and therapeutic effectiveness. Thus, differences between self-report and other informants might be used as a therapeutic tool. Such interpretations can be further advanced by studies that aim at understanding the multiple causes underlying discrepancies between informants (e.g. De Los Reyes, 2009).

Conclusion

Stable differences between informants reflect the stability of unique informant reports, which are well captured by the CT-C(M-1) model. More discrepancies between informants are predictive of more variability in self-reported problems over time, which is mostly in the direction of the parental report. In this study predictive utility was not impressive, but a full evaluation of predictive utility should be based on more outcomes and on comparisons of multiple predictive algorithms. Furthermore, differences between informants can have important diagnostic and therapeutic utility.

Chapter 7

Discussion

In this thesis, I have confronted a few issues concerning measurement, structure and diagnosis of psychopathology. In chapter 2, confirmatory and exploratory factor analysis were used to create a dimensional framework that captured covariance between multiple domains of parent-reported symptoms. Subscales of a questionnaire related to the 'Broader Autism Phenotype' (i.e. BAP-subscales) were integrated into commonly used latent variable models of internalizing and externalizing psychopathology. Some BAP-subscales loaded on the internalizing factor and others on the externalizing factor. Furthermore, a third factor was added to the model to capture covariance between BAP-subscales. In my view, such dimensional latent variable models are preferable to the many arbitrary dichotomizations of DSM-IV. As argued in chapter 3, this does not imply that dimensions are always more valid and useful. Categorical constructs are appropriate in the case of well-established natural categories. Furthermore, within a dimensional framework categories can be introduced for specific purposes. Utility-based categories can be created by adopting useful cut-off criteria. Sometimes these may be used to guide decision-making. Another reason for introducing categories into a dimensional framework is that continuous scores are not very appropriate for communication among experts and between experts and non-experts. For this reason it was suggested that ordinal variables be used with terms adopted from natural language (low, mild, moderate, etc.).

Chapters 4 to 6 dealt with the large differences that are observed between informant-reports of psychological problems. In chapter 4, the 'mixing and matching' approach to analyzing multi-informant data proposed by Kraemer et al. (2003) was extended in order to analyze reports of internalizing and externalizing problems by three informants: self, parent, and teacher. With Principal Component Analysis, these reports were reduced to the four orthogonal components that were expected on the basis of the work of Kraemer et al. (2003). Two of these components were related to differences between informants, and named Perspective (self versus others) and Context (parent versus teacher). The other two components were not related to differences between informants and were named Severity (many versus few problems) and Direction (internalizing versus externalizing problems). The advantage of this pragmatic approach is that multiple informant-reports are captured within the same dimensional structure. Using the approach in research and clinical practice has the advantage that the large differences between informants in the estimation of psychological problems and hypotheses about the reasons for these differences are taken into account. However the 'Kraemer-Noordhof' approach has disadvantages as well. Of specific concern is the interpretation of the bipolar components Context and Perspective. There is not much proof that variance between individuals on these

components is indeed caused by specific individual differences in contexts and perspectives. The components were found in the non-rotated PCA-solution, but in rotated solutions the first three components generally are: self, parent, and teacher. The reason is that within-informant covariance of different scales (e.g. r=.60 for self-reported internalizing and externalizing problems) is generally higher than between-informant covariance of the same scales (e.g. r=.30). In chapter 5 it was found that estimated correlations between the internalizing and externalizing domain were not significantly influenced by sampling biases, but might have been influenced strongly by response and observation biases. The hypothesis that all correlation resulted from these biases could not be rejected. This hypothesis was not supported either, because the distinction between 'biased' and 'objective', or between 'method'[7] and 'trait', could not be adequately made. This recognition lead me to adopt the Correlated Traits – Correlated Methods minus one model (CT-C(M-1); Eid et al., 2003) which does, in spite of its name, not distinguish between 'methods' and 'traits' (Geiser, et al., 2008). Instead a distinction is made between a reference method (e.g. self-report) and all other methods (e.g. parental or teacher report). The CT-C(M-1) model was also used in chapter 6 to study the stability and predictive utility of differences between informants. In this chapter a distinction was made between two types of differences between informants. On the one hand 'informant discrepancies', which were defined as the total amount of non-directional differences in the item-responses given by informants. On the other hand 'informant deviances', which were defined on the basis of the CT-C(M-1) model as the directional deviance (higher or lower score) of an informant from the reference measure.[8] Deviances, but not discrepancies, were found to have a high rank-order stability. Both discrepancies and deviances were predictors of change in self-reported problems during adolescence, but their predictive utility was modest.

In chapter 6 I emphasized that differences between informants can result from problems in reliability or validity. I believe that these are two different interpretations resulting from different hypotheses about what is measured by the informant-reports. In the following I aim to make a clear distinction between the two interpretations. Subsequently, I will propose a three-step approach for the interpretation of multi-informant data in which considerations regarding reliability and validity are used as two different, but not mutually exclusive, frameworks for interpretation of multi-informant data. Finally, I will show how this model can be applied in research and clinical practice.

---

[7] In the context of this discussion the term 'method' can be replaced by 'informant', but in general and also in chapter 5 it covers more than only the issue of informants.

[8] In the literature and also in the earlier chapters of this paper the term 'informant discrepancies' is used to refer to both types of differences. In this discussion I will simply use the term differences between informant-reports.

Reliability

Reliability is a measure of how precise differences in true scores can be estimated on the basis of test results. If two instruments measure the *same* attribute and the correlation between the test results of these instruments is low, then one or both instruments are unreliable. Similarly, if reports by multiple informants measure the *same* attribute, then a low correlation between these reports indicates unreliability.

It is often assumed that informant-reports are multiple measures of the relative position [9] of individuals on the *same* dimension of psychopathology. This interpretation is strongly suggested when using the same name (e.g. internalizing or externalizing) for variables that are based on reports from different informants. This is also the interpretation that is implicit in questions like: "Which informant should be trusted?" or "How can reports of multiple informants be aggregated into a reliable measure?". The general idea can be illustrated as a line (dimension) on which each informant report and each algorithm based on multiple reports (e.g. the mean, the OR-rule, and the AND-rule) is represented by a particular position (see figure 1).

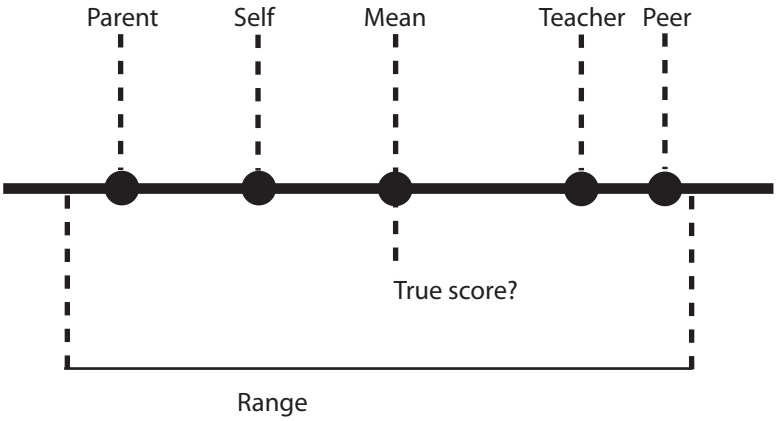

Figure 1. Illustration of the assumption that multiple informant-reports measure the same (dimensional) attribute.

If a large random sample of estimations of the same attribute were available, and these estimations would follow a normal distribution then the mean of the distribution would indicate the true score. However, informant reports of psychopathology are never random and there are no large samples of independent

[9] *Given that there is almost no absolute measurement in psychopathology research, in this discussion I will always assume that scores are indicators of relative positions, like z-scores or T-scores or percentiles.*

informants. Therefore, the mean does not necessarily indicate the true score and I think that the information presented in figure 1 should not be summarized by using a mean score.

Validity

Differences between informants can also be interpreted from the perspective of validity. A concise and clear definition of validity has been given by Borsboom et al. (2004): an instrument is valid for measuring an attribute if and only if (1) the attribute exists, and (2) variance in the attribute produces variance in the test results. The attribute does not necessarily cause all variance in test results.

As discussed in the introduction, the model of internalizing and externalizing psychopathology is a latent variable model capturing the covariance between reported problems in general population samples. The advantage of latent variable models is that multiple alternative representations of the latent structure underlying reported problems can be compared. As was shown in chapters 5 and 6, multi-informant data can also be studied by using latent variable models. Specifically, the CT-C(M-1) model was used to model the structure of multi-informant data (see figure 2). The CT-C(M-1) model can be considered more inclusive than single-informant based latent variable models of psychopathology.

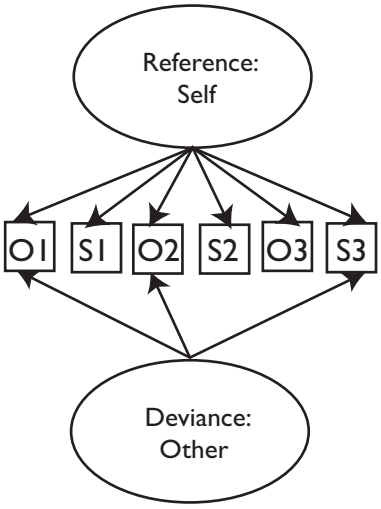

Figure 2. The CT-C(M-1) model with two informants and self-report as reference-method.
Note: O1-O3= observed variables in reports by others. *S1-S3* = observed variables in self-report.

Using CT-C(M-1) as a model of the latent structure underlying reported symptoms has important theoretical implications. If self-report is chosen as the reference method, the estimated 'traits' refer to self perspective. The questions of validity then become: (1) do people have a perspective on their own psychological problems, and (2) does variance in this perspective produce variance in the test-results? Similarly, the 'methods' refer to the deviance of other informants from this self-perspective. Therefore the question is: (1) do other informants have a different perspective, and (2) does variance in this perspective produce variance in the test-results. Obviously, the causal structures underlying ideas about individuals and psychopathology and the translation of these ideas into questionnaire responses are more complicated and less linear than suggested by the simplifications made in the CT-C(M-1) model. Nevertheless, I think that the CT-C(M-1) model provides a useful 'reference model' of the structure underlying reported psychopathology to which alternative latent models of that structure can be compared.

A three-step approach to interpreting multi-informant assessment

Validity and reliability provide two different perspectives on the same observed differences between informant-reports. If informant-reports measure the *same* attribute, then differences indicate unreliability. If informant-reports measure *different* attributes, then differences result from differences between the attributes that are measured. In research and clinical practice both interpretations can be relevant, but they are based on different ideas of what is measured. In the following I will propose a three-step approach for interpreting multi-informant assessment in which a sharp distinction is made between the two interpretations. A crucial, non-arbitrary, choice is that the perspective of an individual on his own psychological problems is used as reference method. The interpretation starts (step 1) with interpreting self-report as the perspective of individuals on their own psychological problems. In the other two steps the reports of other informants are interpreted. On the one hand (step 2) multiple informants are used to estimate the amount of psychological problems along the *same* dimension. In this step it is assumed that differences between informant-reports result from unreliability. On the other hand (step 3) other informants are used to measure attributes that are not measured by self-report. In this step it is assumed that differences between informant-reports result from differences in the attributes that they measure.

In the first step self-report is interpreted as a measure of an individual's own perspective. Completing a questionnaire can be considered as a structured means of communication between an individual and others (researchers, clinicians). The measurement can be regarded valid for measuring individual differences in self-perspective if individuals do indeed differ with regard to their perspective on their own problems, and if these differences *cause* variance in the test results. Validity can also be defined intra-individually: can individuals, on the basis of their own perspective,

influence the test-results in order to indicate which problems they do and do not experience? And in a repeated measures design: do experienced changes in psychological problems cause changes in the test-results? I do not assume that individuals are very adequate and unbiased judges of their own psychological condition, but I have theoretical, pragmatic and ethical arguments for choosing self-report as a reference method. The theoretical argument is that I do not believe that there is a well-established distinction between self-perspective and psychopathology. For most syndromes described in DSM-IV there is no strongly supported theory on which a clear distinction between reported symptoms and underlying psychopathology could be based. This does not imply that self-perspective and psychopathology are synonyms, but that we do not know how to distinguish between subjective experience and objective psychological problems. However, we do know that people choose to consult experts because they are concerned about problems they experience. These private concerns can be used as the starting point or reference method to which new and possibly discrepant sources of information can be compared in order to increase understanding. A pragmatic argument is that self-report is also by far the most commonly used method in research on adult psychopathology. In recent decades many self-report instruments have been developed for children as well. Choosing self-report as a reference method implies consistency with this tradition without accepting self-report as a gold-standard. An ethical argument is that I believe that clinical psychologists and psychiatrists should make attempts at reducing individual psychological suffering. Therefore, the individual experience should be the first and last, not the only, source of information for evaluating whether these attempts have been successful.

The second step is based on the hypothesis that reports of multiple informants refer to the same dimension. Some people are more aggressive or more depressed than others. The hypothesis is that some variance in reports of all relevant informants is caused by individual differences in these attributes. This means that differences between informant-reports are interpreted as *quantitatively* different estimations along the *same* dimension. The true score of individuals on this dimension cannot be determined, because there is no gold-standard nor a large sample of independent estimations [10]. For this reason I would prefer to represent the multiple measures as points on a range. Algorithms could be developed to construct individual ranges for which the chances are small that a new, relevant, informant gives an estimation that is outside of the estimated range. The clinical impression of a therapist could be considered as one point on this range as well. The range constitutes a minimal amount of agreement about an individual's ranking in comparison to others. It is likely that the true score, if it exists, is somewhere on this range.

In the third step self-report is complemented by information that can only be measured with reports by other informants. This means that reports of other

[10] *And it may therefore be questioned whether the 'true score' actually exists.*

informants are considered as *qualitatively* different from self-report: they measure something else. Validity should therefore be evaluated by investigating the specific causes of variance in the reports of these other informants. In chapters 4 to 6 I have mentioned multiple hypothetical causal pathways that may result in differences between informant-reports. Currently most of these pathways have not been rigorously tested, which means that most interpretations of differences between informants are post-hoc arguments that may serve the development of hypotheses rather than the establishment of reliable knowledge. Moreover, it is very well possible that these differences are not homogeneous within the population, but are strongly influenced by idiosyncratic interactions between items, observations, memories, comparisons, judgments and response styles.

Implications for research

This three-step approach can be directly applied in how research is reported. In the results and discussion sections of a scientific report, an explicit distinction can be made between results and interpretations at each step. The step 1 analysis refers to self-reported problems and these are interpreted as reflecting an individual's own perspective[11]. The step 2 analysis refers to a range of estimations based on multiple informants and the informant scores, which are interpreted as reflecting unreliable estimations of the same attribute. The step 3 analysis concerns the deviances between self and other informants, which are interpreted as reflecting differences between the attributes that are measured by the different informants.

I also believe that the approach can be fruitful for developing new lines of research. I think there are important challenges at each step. At the first step the challenge is to find valid measures of the extent to which individuals experience problems and are able and willing to express them. One way to accomplish this is to deepen the understanding of how people judge themselves (Wilson, 2009). Specifically, it may be interesting to develop more understanding of the comparisons people make in natural language in everyday life, which are arguably mostly of an ordinal nature. This involves understanding to what models people compare their own experiences, what kind of stories and memories they retrieve while reporting on a questionnaire and how they use questionnaires to communicate about their own perspective. Interesting examples are some recent studies that have used cognitive interviewing (Presser, et al., 2004) to investigate the validity of self-report questionaires of anxiety sensitivity (Brown, Hawkes, & Tata, 2009) and quality of life (Cremeens, Eiser, & Blades, 2007). Another question is how questionnaire responses are related to life-narratives (McAdams, 2006). An interesting example of combining questionnaire approaches and narrative analysis is given by Lodi-Smith, Geise, Roberts, & Robins (2009). A more practical

[11] *To avoid confusion: I do not imagine this individual perspective as a purely subjective construction detached from other perspectives, observations and objective reality.*

question is whether individuals feel that questionnaires allow them to effectively communicate about their problems and whether this can be improved.

At the second step, the challenge is to find ways in which to characterize a range of estimations. If, during a certain period, a large amount of multi-method repeated measures are collected, ranges can be constructed that include all measurement results. Subsequently, algorithms can be constructed that predict these ranges on the basis of a smaller amount of data. Such a procedure may result in algorithms that provide a reliable range of the estimated relative amount of problems for each individual. The probability that new estimations are outside this range should be small [12], which means that the probability that the "true score" (if it exists at all), is outside this range is also small.

At the third step the challenge is to identify the causes of differences between informants. The same type of research strategies as described for self-report could be used. Furthermore, the 'Attribution Bias Context' model (ABC-model), a theoretical framework described by De Los Reyes and Kazdin (2005), may be used to guide such research. This model is based on a theoretical reflection on the process of observing and judging by informants in the context of clinical assessment.

Implications for clinical practice.

The three-step approach can, to some extent, be followed in psychological assessment in individual cases. First, an assessment is made of an individual's own perspective. Assessors can discuss with clients whether the instruments used give a good representation of their perspective on their problems. If not, alternative self-report instruments should be used and evaluated. Second, the assessor can discuss with the client that self-knowledge is important but has also limitations.. In order to go beyond what a client already knows it will be necessary to find different sources of information. The range of estimations can be used to illustrate that different reports can give different results and that therefore it is not possible to give one precise estimation or diagnosis of psychopathology. This idea can be effectively illustrated by drawing a line as in figure 1 and discussing that different informants can give different estimations on this line. Furthermore, ordinal terms may be used to speak about these ranges (e.g. moderate to severely depressed, low to mildly aggressive, etc.) [13]. This ordinal language is not only closer to natural language, but also less suggestive of precise knowledge. Therefore, I think ordinal measures are preferable to either

[12] There is a trade-off between the chance that a new estimation is outside of the range and the length of the range. With a point-estimate the information is very specific, but the chances are almost perfect that a new estimate will be different. With a very large range the changes are almost perfect that it encompasses all estimates, but the informant-value is very low.

[13] In DSM-IV similar terms are used (e.g. in diagnosis of MDD), but are constructed as more detailed specifications rather than an expression of limited accuracy.

dichotomous (present/absent) or dimensional diagnoses. The judgment of clinicians (e.g. Westen & Shedler, 2007) may also be used as one estimation on the range. Third, assessors can discuss deviances between self-report and reports of other informants, including the assessor him or herself. The aim is to understand the specific reasons for deviances.

A golden step?

An important issue of debate regarding diagnoses op psychopathology is whether a diagnosis should be based on (latent) causes or on (observed) characteristics (Zachar & Kendler, 2007). The three-step approach proposed in the above is mainly related to a descriptive approach. Each step refers to interpretations, rather than conclusions, on the basis of test material. These interpretations can be used to develop hypotheses regarding underlying processes that cause individual psychological suffering. However, in many cases in psychiatry and clinical psychology there do not exist strong tests of these hypotheses.

The situation for many well-known diseases is rather different. This may be illustrated by introducing another step of interpreting multi-informant data. In this step a gold standard is introduced. For example, the test for the presence of the H1N1-virus which causes Swine Flu. There is a very strong believe that this test provides a valid and reliable measure. Such a measure would radically change the meaning of informant reports. In presence of a gold standard, informant reports serve as more or less reliable signals that lead to the detection of the presence or absence of a disease. There is a huge difference between diagnoses made in situations in which a gold standard exists and situations in which such a measure does not exist. I think the difference is so big that one even should better use different words for the two situations in order not to be misunderstood by the public. For example, 'diagnosis' for the situation with a gold standard and 'assessment' for the situation without.

Table 1. *Summary of the three step approach for interpreting multi-informant data and implications for research and clinical practice.*

Step	Interpretation[a]	Self-report	Others	Research goals	Clinician
1	Validity	Measure of own perspective		What instruments are valid and reliable for measuring conscious self-perspective?	'I will first try to understand your own view on what your problems are.'
2	Reliability	Point on a range	Points on same range	How to construct a range at the individual level?	'Together with information from other sources we can make an imprecise estimation of the amount of psychological problems you have in comparison to other people.'
3	Validity	Reference measure	Deviance from reference measure	What causal pathways result in deviances of other informants?	'Now we will try to understand why we found differences between your own views and other sources of formation.'
Gold[b]	Deviance from gold standard	Signal	Signal	What are the best signals to avoid false-positives and false-negatives?	'You may have a disorder. We will need to do test X to ascertain this.'

Note: [a] Interpretation of deviances between different methods of assessment.
[b] Interpretations if there would be a gold standard measure.

Conclusion

Most types of psychological assessment do not result in precise quantitative estimations of psychopathology nor in valid judgments about the presence or absence of disorders. Of course this does not imply that people who seek psychological treatment do not suffer. Some treatments can be effective for reducing this suffering. Furthermore, assessment can be used to develop hypotheses about the causes and consequences underlying individual suffering, which can help to improve treatment. Precise quantitative estimations or diagnoses of psychopathology are not a prerequisite for this trial-and-error process of assessing and treating individual suffering. The three-step approach for multi-informant assessment proposed in this thesis may be a useful tool in this process. I hope this approach may contribute to psychological assessment in research and clinical practice.

Author affiliations

University Medical Center Groningen, dep. Interdisciplinary Center for Psychiatric Epidemiology, Netherlands:
Johan Ormel, PhD.
Albertine J. Oldehinkel, PhD.
Catharina A. Hartman, PhD.

University Medical Center Groningen, dep. Child and Adolescent Psychiatry, Groningen, Netherlands:
Catharina A. Hartman, PhD.

Washington University in St. Louis, Departments of Psychology and Psychiatry, USA:
Robert F. Krueger, PhD.

University of Amsterdam, Department of Clinical Psychology, Netherlands:
Jan Henk Kamphuis, PhD.

Erasmus University , Medical Center, Department of Child & Adolescent Psychiatry, Rotterdam, Netherlands:
Frank C. Verhulst, PhD. MD.
Albertine J. Oldehinkel, PhD.

For questions and comments about this thesis, please contact:
Arjen Noordhof
University of Amsterdam, Department of Clinical Psychology
+31 20 5257028
a.noordhof@uva.nl

Nederlandse samenvatting

Als er geen gouden standaard bestaat

In de kern gaat het proefschrift over de (on)mogelijkheid om objectief vast te stellen of en in welke mate iemand aan een psychische stoornis lijdt: het ontbreken van een gouden standaard. Grofweg behandel ik hierbij twee interessante en gerelateerde vraagstukken uit de psychodiagnostiek. Ten eerste de vraag naar 'de latente structuur van psychopathologie' en ten tweede de vraag naar de meetbaarheid van psychopathologie. In de eerste hoofdstukken gaat het met name om de latente structuur. Latente structuur staat hier tegenover observaties. Uit observaties die gedaan worden willen we uitspraken doen over 'wat het achterliggende probleem is'. Bij bepaalde symptomen hoort (volgens de Diagnostic Statistical Manual, DSM-IV) de diagnose 'Major Depressive Disorder' (MDD) en bij andere symptomen hoort de diagnose 'Generalized Anxiety Disorder' (GAD). De suggestie die dit geeft is dat er twee stoornissen bestaan, MDD en GAD, en dat sommige mensen die stoornis hebben en anderen niet. Dit 'er bestaan twee stoornissen…' is een uitspraak over de latente structuur van psychopathologie. In dit geval een uitspraak die veelvuldig betwist wordt en die ook niet volgt uit een nauwkeurig toepassen van DSM-IV. Het is maar zeer de vraag of iemand bij wie de diagnose MDD en de diagnose GAD zijn vastgesteld, lijdt aan twee psychische stoornissen. Ook voor veel andere diagnoses in DSM-IV zijn zeer weinig aanwijzingen dat ze zouden gaan om 'discrete stoornissen': ziekten die je wel of niet hebt. Een veelvuldig aangeprezen alternatief is om mensen in te schalen op een aantal 'dimensies' ('hoog depressief', 'behoorlijk teruggetrokken', 'matig agressief', etc.). Dit lost op zich het probleem van latente structuur niet direct op, want ook de uitspraak 'er bestaan drie dimensies van psychische problemen' kent vele problemen. Het zorgt echter wel voor een relativering van de pretentie 'ziekten' te diagnosticeren. Daarnaast zijn er methoden om deze dimensies zodanig te kiezen dat symptomen die in de algemene bevolking vaak tegelijkertijd voorkomen, zoals 'piekeren' en 'slapeloosheid', onder dezelfde dimensie vallen [1]. Deze dimensies zijn niet direct observeerbaar en worden daarom latente variabelen genoemd. Een model met verschillende dimensies heet daarom een 'latente variabelen model'. Het samen voorkomen van symptomen wil natuurlijk niet zeggen dat ze uitingen zijn van 'hetzelfde', maar het is plausibel dat wanneer symptomen vaak samengaan ze in een bepaald oorzakelijk verband staan en dus 'met elkaar te maken hebben'. Het vaak samen gaan van symptomen wordt uitgedrukt in hun covariantie en latente modellen worden zodanig gekozen dat ze zoveel mogelijk van deze covariantie 'verklaren'.

[1] *Voor de volledigheid: het is niet zo dat dergelijke modellen alleen met dimensies kunnen worden gemaakt, maar in de praktijk worden meestal dimensies gebruikt en werkt dat goed.*

In het tweede hoofdstuk heb ik[2] een heel aantal van dergelijke latente variabelen modellen gemaakt en getoetst in hoeverre ze passen bij de verzamelde data. In het onderzoek TRAILS (Tracking Adolescents' Individual Lives Survey) wordt een grote groep jongeren (2230) gevolgd sinds het einde van de basisschool. Inmiddels zijn de meesten van de middelbare school af en is de vierde meting (T4) al een heel eind gevorderd. Ikzelf heb meegewerkt aan de derde meting en gebruik in mijn onderzoek vragenlijstgegevens van de eerste drie metingen (T1-T3). Het onderzoek dat ik in het tweede hoofdstuk presenteer is gebaseerd op door ouders gerapporteerde klachten. Specifiek heb ik een bestaand en veel gebruikt dimensioneel model (het model van internalizerende en externalizerende problemen) uitgebreid met klachten die daar tot nu toe niet in opgenomen waren. Dit zijn klachten die veelvuldig voorkomen bij kinderen met de diagnosen 'autistische stoornis', 'Asperger's syndroom' en 'PDD-NOS' [3]. Het bleek goed mogelijk om deze klachten op te nemen in het model, maar daarvoor moest het model worden uitgebreid met een dimensie die specifiek te maken heeft met 'autisme-achtige problemen', in het Engels: 'The broader autism phenotype (BAP)'. Dit resultaat zou in de praktijk kunnen worden toegepast, onder meer door bij kinderen die nu 'PDD-NOS'-ers heten een meer genuanceerde beschrijving van problemen op meerdere dimensies te geven. Overigens betwijfel ik of de in hoofdstuk twee gebruikte term 'autism phenotype' wel zo geschikt zou zijn voor de praktijk, aangezien het woord autisme, bijvoorbeeld in films, zeer specifieke en zelfs stigmatiserende associaties oproept die bij veel van deze kinderen helemaal niet aan de orde zijn.

Het derde hoofdstuk gaat verder in op de praktische toepassing van deze dimensionele modellen in de klinische praktijk. Met name gaat het in dit hoofdstuk om de vraag of het mogelijk en nuttig is om toch categorieën uit dimensionele scores te maken. Ik laat zien hoe dit zou kunnen en bespreek verschillende doelstellingen waarvoor dit nuttig zou zijn. De belangrijkste zijn communicatie (het communiceert niet makkelijk met getallen op 4 dimensies) en het nemen van beslissingen (moet iemand nu wel of niet een diagnose en behandeling krijgen). Wat betreft het communicatieprobleem zijn er bruikbare manieren om van dimensionele scores hanteerbare taal te maken. Het beslissingsprobleem ligt veel ingewikkelder en wordt in hoofdstuk 3 slechts gedeeltelijk opgelost. Enerzijds wordt besproken dat bij veel beslissingen lokale informatie van groot belang is en het dus niet per se optimaal is om hier internationaal vaste criteria voor te schrijven. Anderzijds wordt gesteld dat het voor sommige maatschappelijke doeleinden nodig is dat experts beslissen wie wel en wie niet behandeling (vergoed) zouden moeten krijgen.

[2] *Ik gebruik in deze samenvatting steeds de ik-vorm, maar het werk voor ieder hoofdstuk is steeds een samenwerking met meerdere auteurs geweest.*

[3] *In Nederland wordt meestal deze Engelse afkorting van Pervasive Developmental Disorder – Not Otherwise Specified gebruikt, dus pervasieve ontwikkelingsstoornis – Niet Anderszins Omschreven.*

Voor deze expertbeslissingen zou het verstandig kunnen zijn om algemene criteria vast te leggen en te evalueren in hoeverre goed geïnformeerde experts met elkaar tot overeenstemming kunnen komen. Dit blijft echter een heikel punt, des te meer door de grote verschillen tussen verschillende bronnen van informatie.

De hoofdstukken 4 tot 6 gaan over deze grote verschillen. Het blijkt dat verschillende informanten (leraar, ouder, zelf) die dezelfde vragenlijsten over hetzelfde kind invullen tot heel verschillende rapportages komen. Deze conclusie komt voort uit het observeren van covariantie (zie boven) tussen rapportages van verschillende informanten op dezelfde vragenlijstitems. Het blijkt dat heel verschillende probleemgebieden (bv. agressief en teruggetrokken gedrag) gerapporteerd door eenzelfde informant vaak duidelijk méér covariantie vertonen dan hetzelfde probleemgebied gerapporteerd door verschillende informanten. Het maakt dus voor de schatting van de mate van psychische problemen nogal wat uit aan wie je om informatie vraagt.

In hoofdstuk 4 heb ik een poging ondernomen om de informatie van verschillende informanten (zelf, leraar, ouder) over verschillende probleemgebieden (agressie, teruggetrokken gedrag, angst en meer) te integreren in één dimensioneel model. De cruciale assumptie waarop dit model stoelt is dat de meeste verschillen tussen informanten worden veroorzaakt door systematische verschillen tussen het perspectief van waaruit informanten naar een persoon kijken en de context waarin zij met die persoon te maken krijgen. Informanten zou je moeten selecteren op basis van de mate waarin ze overeenkomen én verschillen wat betreft perspectief en context. Leraar en ouder komen bijvoorbeeld deels overeen wat betreft dit perspectief, maar verschillen juist wat betreft context (school versus thuis). Vandaar dat in dit model de systematische verschillen tussen zelf-rapportage en rapportage door leraar en ouder wordt aangeduid met de term Perspectief, terwijl systematische verschillen tussen leraar en ouder worden aangeduid met de term Context. De zo gevormde componenten Context en Perspectief blijken samen een zeer belangrijk deel van de variantie in de gerapporteerde scores te 'verklaren'. Hieruit blijkt opnieuw de grote invloed van de gekozen informanten op de uiteindelijke meting. Er wordt bovendien de belangrijke stap gezet om na te denken over waarom informanten van elkaar verschillen. Een heel andere manier om over deze verschillen na te denken komt aan de orde in hoofdstuk 5.

In hoofdstuk 5 stel ik enigszins provocerend de vraag aan de orde of covariantie tussen zeer verschillende probleemgebieden (externalizerend: agressie, delinquentie, etc. en internalizerende: teruggetrokken gedrag, angstig, depressief, etc.) wel echt bestaat (fact) of dat het slechts voortkomt uit de methode van meten (artefact). Het is een nogal technisch artikel, waarin ik een vijftal methodische problemen bespreek en laat zien dat met name de gebruikte informanten en instrumenten invloed hebben op het inschatten van de associatie tussen deze verschillende probleemgebieden. Uiteindelijk kon ik de hypothese dat het slechts om een artefact gaat niet verwerpen. Erg waarschijnlijk vind ik het overigens niet dat het 'enkel een artefact' is. In de

discussie bij dit hoofdstuk bespreek ik dat het onderscheid tussen fact en artefact in het geval van structureel verschillende informanten niet goed gemaakt kan worden. Dat wat iemand zelf rapporteert kun je niet simpelweg als 'artefact' beschouwen, maar is eerder het resultaat van een specifieke interactie tussen iemands eigen ervaringen en herinneringen, de aangeboden vragen en de manier waarop iemand besluit tot een bepaalde respons (0, 1 of 2).

Toch blijft het merkwaardig dat verschillende metingen van 'hetzelfde' tot zulke verschillende resultaten leiden. In hoofdstuk 6 stel ik daarom dat we moeten kiezen: ofwel de metingen meten dezelfde dimensie, maar dan meten ze die wel bijzonder onnauwkeurig. Ofwel, de metingen meten verschillende dimensies, maar dan is het eigenlijk vreemd om deze verschillende dimensies dezelfde naam (bijvoorbeeld agressie) te geven. Wat ik verder laat zien in hoofdstuk zes is dat de verschillen tussen zelf-rapportage en rapportage door een ouder behoorlijk stabiel zijn. Dat wil zeggen dat als de ouder op de basisschool meer, of juist minder, rapporteert dan het kind, dit verschil waarschijnlijk gedurende de adolescentie zal blijven. Wel is het zo dat zelf-rapportage gemiddeld enigszins verandert in de richting van ouder-rapportage: als een ouder meer problemen rapporteert dan is er een grotere kans dat zelf-rapportage zal toenemen. Dit effect is echter klein en het is maar de vraag of dit gegeven erg bruikbaar is in de klinische praktijk. De vuistregel is dat je kunt verwachten dat verschillen blijven bestaan.

Dit betekent dat we bij twee informanten te maken hebben met twee behoorlijk stabiele, maar toch zeer verschillende schattingen van psychische problemen. In het laatste hoofdstuk heb ik een poging gewaagd om uit deze resultaten te komen tot een manier om met informatie van meerdere informanten om te gaan wanneer er geen gouden standaard bestaat. Twee keuzes zijn hierbij cruciaal. Ten eerste kies ik ervoor om zelf-rapportage als een referentiemethode te beschouwen. Een referentiemethode is in zekere zin het tegenovergestelde van een gouden standaard. Waar een gouden standaard het onbetwistbare eindpunt van twijfel en discussie markeert, is een referentiemethode juist het betwistbare beginpunt daarvan. Betwistbaar, omdat ik er niet vanuit ga dat iemand zelf een uitermate goede beoordelaar is van zijn of haar eigen psychologische conditie. Beginpunt, omdat de zelf gevoelde symptomen het punt van vertrek en de motivatie (zouden moeten) zijn voor het starten van hulpverlening. De bewuste eigen ervaring is dus het (onvervangbare) beginpunt waartegen nieuwe perspectieven van andere informanten en van behandelaars kunnen worden afgezet. Stap 1 in mijn voorstel betreft daarom de interpretatie van zelf-rapportage als zelf-rapportage, en dus niet als directe meting van pathologie. De tweede keuze is dat ik een expliciet onderscheid maak tussen een kwantitatieve en een kwalitatieve interpretatie van verschillen tussen zelf-rapportage en rapportage door andere informanten. Dat wil zeggen: ofwel andere informanten meten hetzelfde, maar de metingen verschillen *kwantitatief*, dat wil zeggen ze zijn niet betrouwbaar. Ofwel andere informanten meten *kwalitatief* iets anders. In stap 2 van mijn voorstel worden verschillen kwantitatief geïnterpreteerd. Dit komt er met name op neer dat de

onnauwkeurigheid van meting expliciet wordt in de rapportage. Een manier waarop dit zou kunnen is door in plaats van diagnoses (bv. ziekte wel/niet aanwezig) of exacte scores (bv. neuroticisme = 9) gebruik te maken van een interval aangeduid met onnauwkeurige termen uit de natuurlijke taal (bv. 'mild tot behoorlijk depressief', 'laag tot gemiddeld angstig'). In stap 3 worden verschillen kwalitatief geïnterpreteerd. Ik geloof echter niet dat er op dit moment duidelijke interpretatieregels hiervoor beschikbaar zijn. Dat wil zeggen dat deze interpretatie in de klinische praktijk neer zou komen op een gesprek waarin wordt gepoogd te begrijpen hoe verschillen in het individuele geval tot stand zijn gekomen.

References

Achenbach, T. M. (1966). Classification of children's psychiatric symptoms - a factor-analytic study. *Psychological Monographs, 80*(7).

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

Achenbach, T. M. (1991b). *Manual for the Teacher Rating Form and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991c). *Manual for the Youth Self-Report and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Becker, A., Dopfner, M., Heiervang, E., Roessner, V., Steinhausen, H. C., et al. (2008). Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications, and future directions. *Journal of Child Psychology and Psychiatry, 49*(3), 251-275.

Achenbach, T. M., & Edelbrock, C. S. (1978). Classification of child psychopathology - review and analysis of empirical efforts. *Psychological Bulletin, 85*(6), 1275-1301.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child Adolescent Behavioral and Emotional-Problems - Implications of Cross-Informant Correlations for Situational Specificity. *Psychological Bulletin, 101*(2), 213-232.

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

Baldessarini, R. J., Finklestein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry, 40*, 569-573.

Bishop, D. V. M., & Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication: use of the Children's Communication Checklist in a clinical setting. *Developmental Medicine and Child Neurology, 43*(12), 809-818.

Bolton, P., Macdonald, H., Pickles, A., Rios, P., Goode, S., Crowson, M., et al. (1994). A case-control family history of autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 35*(5), 877-900.

Boomsma, A., Van Lang, N. D. J., De Jonge, M. V., De Bildt, A. A., Van Engeland, H., & Minderaa, R. B. (2008). A new symptom model for autism cross-validated in an independent sample. *Journal of Child Psychology and Psychiatry, 49*(8), 809-816.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Brereton, A. V., Tonge, B. J., & Einfeld, S. L. (2006). Psychopathology in children and adolescents with autism compared to young people with intellectual disability. *Journal of Autism and Developmental Disorders, 36*(7), 863-870.

Brown, G. P., Hawkes, N. C., & Tata, P. (2009). Construct validity and vulnerability to anxiety: A cognitive interviewing study of the revised Anxiety Sensitivity Index. *Journal of Anxiety Disorders, 23*(7), 942-949.

Burt, S.A., McGue, M., Krueger, R. F., & Iacono, W. G. (2005). Sources of covariation among the child-externalizing disorders: informant effects and the shared environment. *Psychological Medicine, 35*(8), 1133-1144.

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. *Behaviour Research and Therapy, 38*(8), 835-855.

Clark, L. A. (2005). Temperament as a unifying basis for personality and psychopathology. *Journal of Abnormal Psychology, 114*(4), 505-521.

Clark, L. A., & Harrison, D. (2001). Assessment instruments. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research and treatment.* . New York: Guilford Press.

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 45*, 719-726.

Constantino, J. N., & Todd, R. D. (2003). Autistic traits in the general population - A twin study. *Archives of General Psychiatry, 60*(5), 524-530.

Courvoisier, D. S., Nussbeck, F. W., Eid, M., & Cole, D. A. (2008). Analyzing the convergent and discriminant validity of states and traits: Development and applications of multimethod latent state-trait models. *Psychological Assessment, 20*(3), 270-280.

Cremeens, J., Eiser, C., & Blades, M. (2007). A qualitative investigation of school-aged children's answers to items from a generic quality of life measure. *Child Care Health and Development, 33*(1), 83-89.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

De Bruin, E. I., Ferdinand, R. F., Meester, S., de Nijs, P. F. A., & Verheij, F. (2007). High rates of psychiatric co-morbidity in PDD-NOS. *Journal of Autism and Developmental Disorders, 37*(5), 877-886.

De Groot, A., Koot, H., & Verhulst, F. (1994). Cross-cultural generalizability of the Child Behavior Checklist cross-informant syndromes. *Psychological Assessment, 6*(3), 225-230.

De Jonge, P., & Slaets, J. P. J. (2005). Response sets in self-report data and their associations with personality traits. *European Journal of Psychiatry, 19*(4), 209-214.

De Los Reyes, A. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology, 37 (5),* 637-652.

De Los Reyes, A., Goodman, K. L., Kliewer, W., & Reid-Quinones, K. (2008). Whose depression relates to discrepancies? Testing relations between informant characteristics and informant discrepancies from both informants' perspectives. *Psychological Assessment, 20*(2), 139-149.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*(4), 483-509.

De Los Reyes, A., & Kazdin, A. E. (2006). Conceptualizing changes in behavior in intervention research: the range of possible changes model. *Psychological Review, 113*(3), 554-583.

De Winter, A., Oldehinkel, A. J., Veenstra, R., Brunnekreef, J. A., Verhulst, F. C., & Ormel, J. (2005). Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. *European Journal of Epidemiology, 20*(2), 173-181.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8*(1), 38-60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*(3), 230-253.

Essex, M. J., Klein, M. H., Cho, E., & Kraemer, H. C. (2003). Exposure to maternal depression and marital conflict: Gender differences in children's later mental health symptoms. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(6), 728-737.

Essex, M. J., Kraemer, H. C., Armstrong, J. M., Boyce, T., Goldsmith, H. H., Klein, M. H., et al. (2006). Exploring risk factors for the emergence of children's mental health problems. *Archives of General Psychiatry, 63*(11), 1246-1256.

Feinberg, M. E., Howe, G. W., Reiss, D., & Hetherington, E. M. (2000). Relationship between perceptual differences of parenting and adolescent antisocial behavior and depressive symptoms. *Journal of Family Psychology, 14*(4), 531-555.

Ferdinand, R. F., van der Ende, J., & Verhulst, F. C. (2004). Parent-adolescent disagreement regarding psychopathology in adolescents from the general population as a risk factor for adverse outcome. *Journal of Abnormal Psychology, 113*(2), 198-206.

Ferdinand, R. F., Van Der Ende, J., & Verhulst, F. C. (2006). Prognostic value of parent-adolescent disagreement in a referred sample. *European Child & Adolescent Psychiatry, 15*(3), 156-162.

Finn, S. E. (1982). Base rates, utilities, and DSM-II - Schortcomings of fixed-rule systems of psychodiagnosis. *Journal of Abnormal Psychology, 91*(4), 294-302.

Finn, S. E. (1983). Utility-balanced and utility-imbalanced rules - reply. *Journal of Abnormal Psychology, 92*(4), 499-501.

Folstein, S. E., & Rutter, M. L. (1988). Autism - familial aggregation and genetic implications. *Journal of Autism and Developmental Disorders, 18*(1), 3-30.

Frances, A. J. (1993). Dimensional diagnosis of personality: Not whether, but when and which. *Psychological Inquiry, 4*(2), 110-111.

Frances, A. J., First, M. B., Widiger, T. A., Miele, G. M., Tilly, S. M., Davis, W. W., et al. (1991). An A to Z guide to DSM-IV conundrums. *Journal of Abnormal Psychology, 100*(3), 407-412.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods, 13*(1), 49-57.

Ghaziuddin, M., Ghaziuddin, N., & Greden, J. (2002). Depression in persons with autism: Implications for research and clinical care. *Journal of Autism and Developmental Disorders, 32*(4), 299-306.

Gilmour, J., Hill, B., Place, M., & Skuse, D. H. (2004). Social communication deficits in conduct disorder: a clinical and community survey. *Journal of Child Psychology and Psychiatry, 45*(5), 967-978.

Gomarus, H. K., Wijers, A. A., Minderaa, R. B., & Althaus, M. (2009). ERP correlates of selective attention and working memory capacities in children with ADHD and/or PDD-NOS. *Clinical Neurophysiology, 120*(1), 60-72.

Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition & Emotion, 4*(3), 269-288.

Grimm, K. J., Pianta, R. C., & Konold, T. (2009). Longitudinal Multitrait-Multimethod Models for Developmental Research. *Multivariate Behavioral Research, 44*(2), 233-258.

Grove, W. M. (1985). Bootstrapping diagnoses using bayes theorem - its not worth the trouble. *Journal of Consulting and Clinical Psychology, 53*(2), 261-263.

Guion, K., Mrug, S., & Windle, M. (2009). Predictive Value of Informant Discrepancies in Reports of Parenting: Relations to Early Adolescents' Adjustment. *Journal of Abnormal Child Psychology, 37*(1), 17-30.

Happe, F., & Ronald, A. (2008). The 'Fractionable Autism Triad': A Review of Evidence from Behavioural, Genetic, Cognitive and Neural Research. *Neuropsychology Review, 18*(4), 287-304.

Harkness, A. R., Tellegen, A., & Waller, N. (1995). Differential Convergence of Self-Report and Informant Data for Multidimensional Personality Questionnaire Traits - Implications for the Construct of Negative Emotionality. *Journal of Personality Assessment, 64*(1), 185-204.

Hartman, C. A., Hox, J., Auerbach, J., Erol, N., Fonseca, A. C., Mellenbergh, G. J., et al. (1999). Syndrome dimensions of the Child Behavior Checklist and the Teacher Report Form: a critical empirical evaluation. *Journal of Child Psychology and Psychiatry, 40*(7), 1095-1116.

Hartman, C. A., Luteijn, E., Serra, M., & Minderaa, R. (2006). Refinement of the children's social behavior questionnaire (CSBQ): An instrument that describes the diverse problems seen in milder forms of PDD. *Journal of Autism and Developmental Disorders, 36*(3), 325-342.

Haslam, N. (2003). Categorical versus dimensional models of mental disorder: the taxometric evidence. *Australian and New Zealand Journal of Psychiatry, 37*(6), 696-704.

Helzer, J. E., Kraemer, H. C., & Krueger, R. F. (2006). The feasibility and need for dimensional psychiatric diagnoses. *Psychological Medicine, 36*(12), 1671-1680.

Helzer, J. E., van den Brink, W., & Guth, S. E. (2006). Should there be both categorical and dimensional criteria for the substance use disorders in DSM-V? *Addiction, 101*, 17-22.

Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the autism-spectrum quotient (AQ): A study in dutch population and patient groups. *Journal of Autism and Developmental Disorders, 38*(8), 1555-1566.

Hoekstra, R. A., Bartels, M., Hudziak, J. J., van Beijsterveldt, T., & Boomsma, D. I. (2007). Genetic and environmental covariation between autistic traits and Behavioral problems. *Twin Research and Human Genetics, 10*(6), 853-860.

Hofler, M. (2005). The effect of misclassification on the estimation of association: a review. *International Journal of Methods in Psychiatric Research, 14*(2), 92-101.

Hofler, M., Lieb, R., & Wittchen, H. U. (2007). Estimating causal effects from observational data with a model for multiple bias. *International Journal of Methods in Psychiatric Research, 16*(2), 77-87.

Hsu, L. M. (1988). Fixed versus flexible MMPI diagnostic rules. *Journal of Consulting and Clinical Psychology, 56*(3), 458-462.

Huisman, M., Oldehinkel, A. J., de Winter, A., Minderaa, R. B., de Bildt, A., Huizink, A. C., et al. (2008). Cohort Profile: The Dutch TRacking Adolescents Individual Lives Survey; TRAILS. *International Journal of Epidemiology, 37*(6), 1227-1235.

Kamphuis, J. H., Finn, S. E., & Butcher, J. N. (2002). Incorporating base rate information in daily clinical decision making. *Clinical personality assessment: Practical approaches (2nd ed.).* (pp. 256-268). New York, NY US: Oxford University Press.

Keiley, M. K., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2000). A cross-domain growth analysis: Externalizing and internalizing behaviors during 8 years of childhood. *Journal of Abnormal Child Psychology, 28*(2), 161-179.

Kendell, R., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry, 160*(1), 4-12.

Kessler, R. C. (2002). Epidemiological perspectives for the development of future diagnostic systems. *Psychopathology, 35*(2-3), 158-161.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 617-627.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry, 160*(9), 1566-1577.

Kraemer, H. C., Noda, A., & O'Hara, R. (2004). Categorical versus dimensional approaches to diagnosis: methodological challenges. *Journal of Psychiatric Research, 38*(1), 17-25.

Kraemer, H. C., Wilson, K. A., & Hayward, C. (2006). Lifetime prevalence and pseudocomorbidity in psychiatric research. *Archives of General Psychiatry, 63*(6), 604-608.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*(10), 921-926.

Krueger, R. F., Caspi, A., Moffitt, T. E., & Silva, P. A. (1998). The structure and stability of common mental disorders (DSM-III-R): A longitudinal-epidemiological study. *Journal of Abnormal Psychology, 107*(2), 216-227.

Krueger, R. F., Chentsova-Dutton, Y. E., Markon, K. E., Goldberg, D., & Ormel, J. (2003). A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *Journal of Abnormal Psychology, 112*(3), 437-447.

Krueger, R. F., & Markon, K. E. (2006a). Reinterpreting comorbidity: a model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2*, 111-133.

Krueger, R. F., & Markon, K. E. (2006b). Understanding psychopathology: Melding behavior genetics, personality, and quantitative psychology to develop an empirically based model. *Current Directions in Psychological Science, 15*(3), 113-117.

Kubarych, T. S., Aggen, S. H., Hettema, J. M., Kendler, K. S., & Neale, M. C. (2005). Endorsement frequencies and factor structure of DSM-III-R and DSM-IV Generalized Anxiety Disorder symptoms in women: implications for future research, classification, clinical practice and comorbidity. *International Journal of Methods in Psychiatric Research, 14*(2), 69-81.

Lahey, B. B., Rathouz, P. J., Van Hulle, C., Urbano, R. C., Krueger, R. F., Applegate, B., et al. (2008). Testing structural models of DSM-IV symptoms of common forms of child and adolescent psychopathology. *Journal of Abnormal Child Psychology, 36*(2), 187-206.

Lilienfeld, S. O. (2003). Comorbidity between and within childhood externalizing and internalizing disorders: Reflections and directions. *Journal of Abnormal Child Psychology, 31*(3), 285-291.

Lilienfeld, S. O., Waldman, I. D., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice, 1*(1), 71-83.

Lodi-Smith, J., Geise, A. C., Roberts, B. W., & Robins, R. W. (2009). Narrating Personality Change. *Journal of Personality and Social Psychology, 96*(3), 679-689.

Lubke, G., & Muthen, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21-39.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41*(4), 499-532.

Luteijn, E., Jackson, S. A. E., Volkmar, F. R., & Minderaa, R. B. (1998). Brief report: The development of the Children's Social Behavior Questionnaire: Preliminary data. *Journal of Autism and Developmental Disorders, 28*(6), 559-565.

Luteijn, E., Luteijn, F., Jackson, S., Volkmar, F., & Minderaa, R. (2000). The children's social behavior questionnaire for milder variants of PDD problems: Evaluation of the psychometric characteristics. *Journal of Autism and Developmental Disorders, 30*(4), 317-330.

Mandy, W. P. L., & Skuse, D. H. (2008). Research Review: What is the association between the social-communication element of autism and repetitive interests, behaviours and activities? *Journal of Child Psychology and Psychiatry, 49*(8), 795-808.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory Factor-Analysis of Multitrait-Multimethod Self-Concept Data - Between-Group and Within-Group Invariance Constraints. *Multivariate Behavioral Research, 28*(3), 313-349.

Matthys, W., Cuperus, J. M., & Van Engeland, M. (1999). Deficient social problem-solving in boys with ODD/CD, with ADHD, and with both disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*(3), 311-321.

McAdams, D. P. (2006). The role of narrative in personality psychology today. *Narrative Inquiry, 16*(1), 11-18.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241.

McGrath, R. E. (2008). Predictor combination in binary decision-making situations. *Psychological Assessment, 20*(3), 195-205.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60*(1), 117-174.

Meehl, P. E. (2001). Comorbidity and taxometrics. *Clinical Psychology: Science and Practice, 8*(4), 507-519.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin, 52*(3), 194-216.

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49*, 377-412.

Moffitt, T. E., & Silva, P. A. (1988). Self-reported delinquency - results from an instrument for New-Zealand. *Australian and New Zealand Journal of Criminology, 21*(4), 227-240.

Mulligan, A., Anney, R. J. L., O'Regan, M., Chen, W., Butler, L., Fitzgerald, M., et al. (2009). Autism symptoms in Attention-Deficit/Hyperactivity Disorder: A Familial trait which Correlates with Conduct, Oppositional Defiant, Language and Motor Disorders. *Journal of Autism and Developmental Disorders, 39*(2), 197-209.

Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance of screening and diagnostic-tests - application of receiver operating characteristic analysis. *Archives of General Psychiatry, 44*(6), 550-555.

Neale, M. C., & Kendler, K. S. (1995). Models of comorbidity for multifactorial disorders. *American Journal of Human Genetics, 57*(4), 935-953.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., et al. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(8), 1078-1085.

Oldehinkel, A. J., Hartman, C. A., De Winter, A. F., Veenstra, R., & Ormel, J. (2004). Temperament profiles associated with internalizing and externalizing problems in preadolescence. *Development and Psychopathology, 16*(2), 421-440.

Ormel, J., Oldehinkel, A. J., Ferdinand, R. F., Hartman, C. A., De Winter, A. F., Veenstra, R., et al. (2005). Internalizing and externalizing problems in adolescence: general and dimension-specific effects of familial loadings and preadolescent temperament traits. *Psychological Medicine, 35*(12), 1825-1835.

Patrick, C. J., Hicks, B. M., Nichol, P. E., & Krueger, R. F. (2007). A bifactor approach to modeling the structure of the psychopathy checklist-revised. *Journal of Personality Disorders, 21*(2), 118-141.

Pelton, J., & Forehand, R. (2001). Discrepancy between mother and child perceptions of their relationship: I. Consequences for adolescents considered within the context of parental divorce. *Journal of Family Violence, 16*(1), 1-15.

Pelton, J., Steele, R. G., Chance, M. W., Forehand, R., & Family Hlth Project Res, G. (2001). Discrepancy between mother and child perceptions of their relationship: II. Consequences for children considered within the context of maternal physical illness. *Journal of Family Violence, 16*(1), 17-35.

Perren, S., Von Wyl, A., Stadelmann, S., Burgin, D., & Von Klitzing, K. (2006). Associations between behavioral/emotional difficulties in kindergarten children and the quality of their peer relationships. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(7), 867-876.

Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and Psychopathology, 15*(3), 529-551.

Pine, D. S., Guyer, A. E., Goldwin, M., Towbin, K. A., & Leibenluft, E. (2008). Autism spectrum disorder scale scores in pediatric mood and anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*(6), 652-661.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., et al. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109-130.

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. C., et al. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders - Implications for research and policy. *Archives of General Psychiatry, 55*(2), 109-115.

Reiersen, A. M., Constantino, J. N., Grimmer, M., Martin, N. G., & Todd, R. D. (2008). Evidence for Shared Genetic Influences on Self-Reported ADHD and Autistic Symptoms in Young Adult Australian Twins. *Twin Research and Human Genetics, 11*(6), 579-585.

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness - its application to schizophrenia. *American Journal of Psychiatry, 126*(7), 983-987.

Ronald, A., Simonoff, E., Kuntsi, J., Asherson, P., & Plomin, R. (2008). Evidence for overlapping genetic influences on autistic and ADHD behaviours in a community twin sample. *Journal of Child Psychology and Psychiatry, 49*(5), 535-542.

Ruscio, J. (2009). Assigning Cases to Groups Using Taxometric Results An Empirical Comparison of Classification Techniques. *Assessment, 16*(1), 55-70.

Samuel, D. B., & Widiger, T. A. (2006). Clinicians' judgments of clinical utility: A comparison of the DSM-IV and five-factor models. *Journal of Abnormal Psychology, 115*(2), 298-308.

Shedler, J., & Westen, D. (2004). Refining personality disorder diagnosis: Integrating science and practice. *American Journal of Psychiatry, 161*(8), 1350-1365.

Shiner, R., & Caspi, A. (2003). Personality differences in childhood and adolescence: measurement development, and consequences. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 44*(1), 2-32.

Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., & Baird, G. (2008). Psychiatric disorders in children with autism spectrum disorders: Prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*(8), 921-929.

Spitzer, R. L. (1983). Psychiatric diagnosis - are clinicians still necessary. *Comprehensive Psychiatry, 24*(5), 399-411.

Spitzer, R. L., First, M. B., Shedler, J., Westen, D., & Skodol, A. E. (2008). Clinical utility of five dimensional systems for personality diagnosis - A "Consumer Preference" study. *Journal of Nervous and Mental Disease, 196*(5), 356-374.

Strauss, M. E., & Smith, G. T. (2009). Construct Validity: Advances in Theory and Methodology. *Annual Review of Clinical Psychology, 5*, 1-25.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*(3), 209-219.

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

Towbin, K. E., Pradella, A., Gorrindo, T., Pine, D. S., & Leibenluft, E. (2005). Autism spectrum traits in children with mood and anxiety disorders. *Journal of Child and Adolescent Psychopharmacology, 15*(3), 452-464.

Van der Maas, H. L., Dolan, C.V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842-861.

Van der Valk, J. C., van den Oord, E., Verhulst, F. C., & Boomsma, D. I. (2001). Using parental ratings to study the etiology of 3-year-old twins' problem behaviors: Different views or rater bias? *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(7), 921-931.

Verheul, R. (2005). Clinical Utility of Dimensional Models for Personality Pathology. *Journal of Personality Disorders, 19*(3), 283-302.

Verhulst, F. C., van der Ende, J., & Koot, H. (1996). *Handleiding voor de CBCL/4-18. [Manual for the CBCL/4-18]*. Rotterdam: Afdeling Kinder- en jeugdpsychiatrie, Sophia Kinderziekenhuis/Academisch Ziekenhuis Rotterdam/Erasmus.

Verhulst, F. C., van der Ende, J., & Koot, H. (1997a). *Handleiding voor de Teacher's Report Form (TRF)/ [Manual for the Teacher's Report Form (TRF)]*. Rotterdam: Afdeling Kinder- en jeugdpsychiatrie, Sophia Kinderziekenhuis/Academisch Ziekenhuis Rotterdam/Erasmus.

Verhulst, F. C., van der Ende, J., & Koot, H. (1997b). *Handleiding voor de Youth Self-Report (YSR) [Manual for the Youth Self-Report]*. Rotterdam: Afdeling Kinder- en jeugdpsychiatrie, Sophia Kinderziekenhuis/Academisch Ziekenhuis Rotterdam/ Erasmus.

Volk, H. E., Henderson, C., Neuman, R. J., & Todd, R. D. (2006). Validation of population-based ADHD subtypes and identification of three clinically impaired subtypes. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 141B*(3), 312-318.

Volkmar, F. R., Cicchetti, D.V., Dykens, E., Sparrow, S. S., Leckman, J. F., & Cohen, D. J. (1988). An evaluation of the autism behavior checklist. *Journal of Autism and Developmental Disorders, 18*(1), 81-97.

Vollebergh, W. A. M., Iedema, J., Bijl, R.V., de Graaf, R., Smit, F., & Ormel, J. (2001). The structure and stability of common mental disorders - The NEMESIS Study. *Archives of General Psychiatry, 58*(6), 597-603.

Wakefield, J. C., & First, M. B. (2003). Confronting the disticntion between disorder and nondisorder: Confronting the overdiagnosis (False Positives) in DSM-V. . In K. A. Philips, M. B. First & H. A. Pincus (Eds.), *Advancing DSM: Dilemmas in Psychiatric Diagnosis*. Washington: American Psychiatric Association.

Waller, N. G. (2006). Carving nature at its joints: Paul Meehl's development of taxometrics. *Journal of Abnormal Psychology, 115*(2), 210-215.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA US: Sage Publications, Inc.

Watson, D. (2005). Rethinking the mood and anxiety disorders: A quantitative hierarchical model for DSM-V. *Journal of Abnormal Psychology, 114*(4), 522-536.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology, 76*(5), 820-838.

Weiss, B., Susser, K., & Catron, T. (1998). Common and specific features of childhood psychopathology. *Journal of Abnormal Psychology, 107*(1), 118-127.

Westen, D., & Shedler, J. (2007). Personality diagnosis with the Shedler-Westen Assessment Procedure (SWAP): integrating clinical and statistical measurement and prediction. *Journal of Abnormal Psychology, 116*(4), 810-822.

Widiger, T. A. (1983). Utilities and fixed diagnostic rules - comments. *Journal of Abnormal Psychology, 92*(4), 495-498.

Widiger, T. A. (1992). Categorical versus dimensional classification - implications from and for research. *Journal of Personality Disorders, 6*(4), 287-300.

Widiger, T. A., & Clark, L. A. (2000). Toward DSM-V and the classification of psychopathology. *Psychological Bulletin, 126*(6), 946-963.

Wilson, T. D. (2009). Know Thyself. *Perspectives on Psychological Science, 4*(4), 384-389.

Wing, L., & Gould, J. (1979). Severe impairments of social-interaction and associated abnormalities in children - epidemiology and classification. *Journal of Autism and Developmental Disorders, 9*(1), 11-29.

Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Who are the comorbid adolescents? Agreement between psychiatric diagnosis, youth, parent, and teacher report. *Journal of Abnormal Child Psychology, 31*(3), 231-245.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*(2), 113-128.

Zachar, P., & Kendler, K. S. (2007). Psychiatric disorders: A conceptual taxonomy. *American Journal of Psychiatry, 164*(4), 557-565.

Acknowledgment / Dankwoord

First some words in English. I want to thank Robert Krueger and Nicholas Eaton for the great period I spent at their lab in St. Louis, Missouri. I felt really welcome and I am very happy with the results of our collaboration. Also, I want to thank Andres de los Reyes for his inspiration on the topic of informant discrepancies.

En nu verder in het Nederlands. Dit proefschrift is het verslag van een zoektocht vol dwalingen, gissingen en zijpaden. Om het vol te houden heb ik de steun van collega's, vrienden en familie hard nodig gehad.

Mijn promotoren, Hans Ormel en Tineke Oldehinkel, hebben met veel geduld mijn gedachtekronkels gevolgd. Middels kritisch lezen en doorvragen hebben zij vele pogingen gewaagd om mij te dwingen tot helder schrijven en denken. Voor zover dat gelukt is heb ik veel aan hen te danken. Ik heb van hen veel stimulatie en betrokkenheid ervaren bij het tot stand komen van dit proefschrift. Ook Catharina Hartman en Jan Henk Kamphuis hebben hieraan een belangrijke bijdrage geleverd.

Ook wil ik de leden van de beoordelingscommissie bedanken voor hun bijdrage: Wilma Vollebergh, Hans Koot en Rob Meijer.

Gedurende mijn vier jaar in Groningen heb ik met veel fijne collega's samengewerkt. Ik begon gelijktijdig met Esther Bouma en Nienke Bosch. We hebben van alles meegemaakt in deze jaren en ik bewaar veel goede herinneringen aan hen. Fijn dat zij de functie van paranimf wilden vervullen. Ook de andere deelnemers aan TRAILS waren prettige collega's: Harriëtte Riese, Andrea de Winter, Frank Verhulst, Martijn Huisman, Martin Bakker, Kennedy Amone, Andrea Prince, Hanneke Wigman, en vele anderen.

Bij velen die hier genoemd zijn heb ik inspiratie opgedaan voor het denken over wetenschap, methodologie en klinische psychologie. Specifiek wil ik daar nog Henk-Jan Conradi, Henk Kiers en Denny Borsboom aan toevoegen, die door hun eigen wijze van denken en bevragen mijn werk soms in een onverwachte richting hebben gestuurd. Verder wil ik Henk Hallie bedanken voor het in orde brengen van al mijn artikelen in reference manager, waardoor het maken van de referentielijst uiteindelijk niet veel meer was dan een druk op de knop. En voor vele regelzaken met betrekking tot mijn proefschrift wil ik Liesbeth Lindeboom en Martha Meschendorp van harte bedanken.

Gelukkig ben ik tijdens het schrijven van mijn proefschrift getrouwd met Anne-Marie Noordhof-Reijnders. Zij heeft op een aantal momenten een cruciale rol gespeeld en heeft mij met pragmatisme, optimisme en diepe betrokkenheid bijgestaan. Op iets meer afstand geldt dat ook voor veel vrienden en familie. Joost Kerklaan, die tijdens vele wandelingen een luisterend oor bood. Melle van den Berg, Henk-Jan van Alphen, David Hollanders en mijn broertje Christiaan Noordhof, bij wie ik altijd een thuis vond in Amsterdam. En tenslotte mijn ouders, Hilde en Wytse Noordhof, die mij hebben grootgebracht in een omgeving van creativiteit, kritisch denken en intellectuele nieuwsgierigheid.