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.

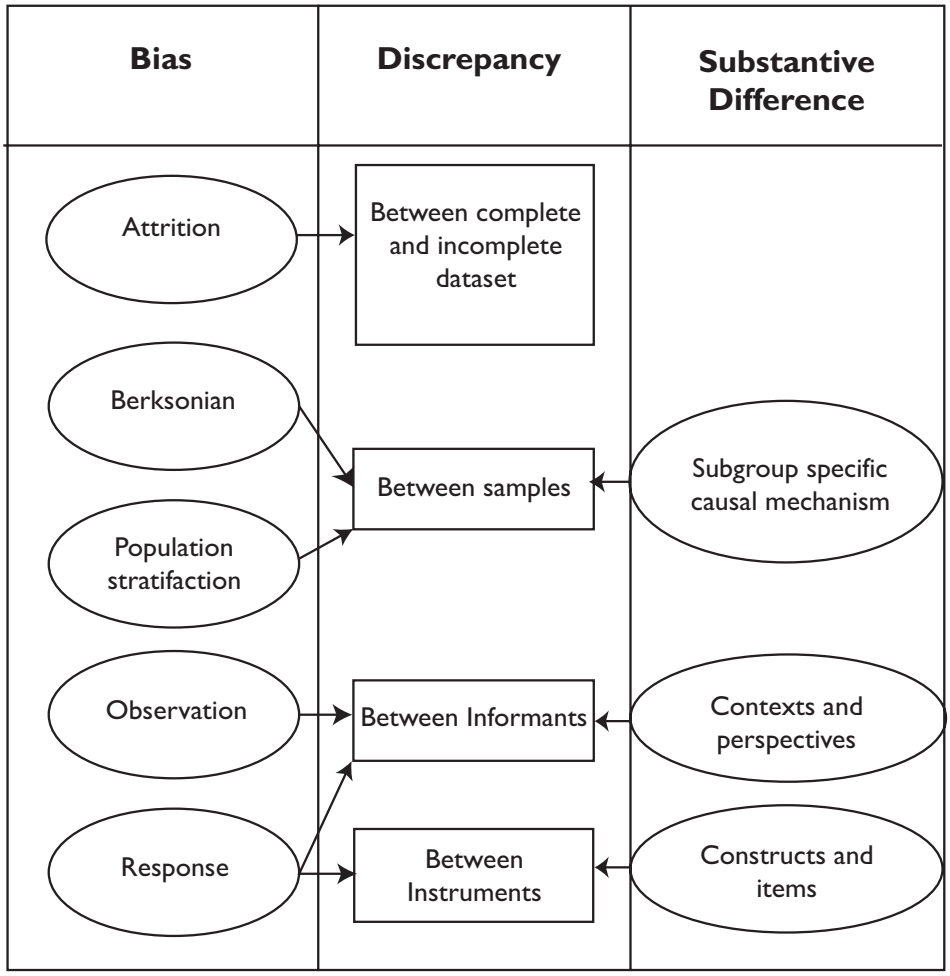

Figure 1. Causes of discrepancy between samples, informants, and instruments.

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.

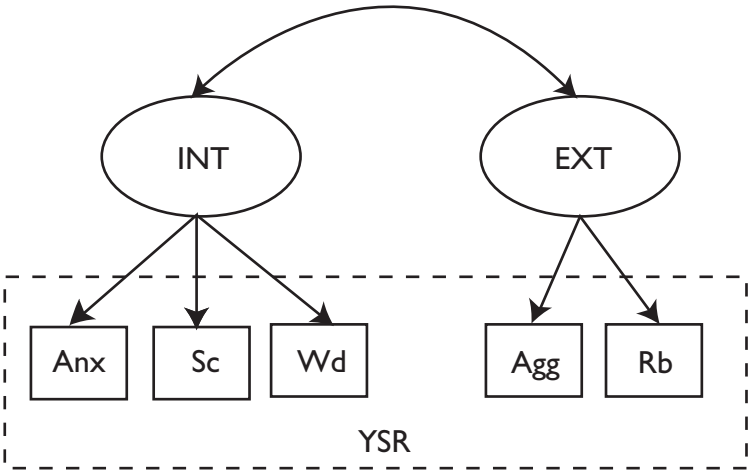

Figure 2. Reference model of correlated Internalizing and Externalizing problems.
Note: Anx = Anxious-Depressed; *Sc* = Somatic Complaints; *Wd* = Withdrawn-Depressed, *Agg* = Aggressive Behavior, *Rb* = Rule-Breaking Behavior, *INT* = Internalizing; *EXT* = Externalizing; *YSR* = Youth Self Report.

• *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.

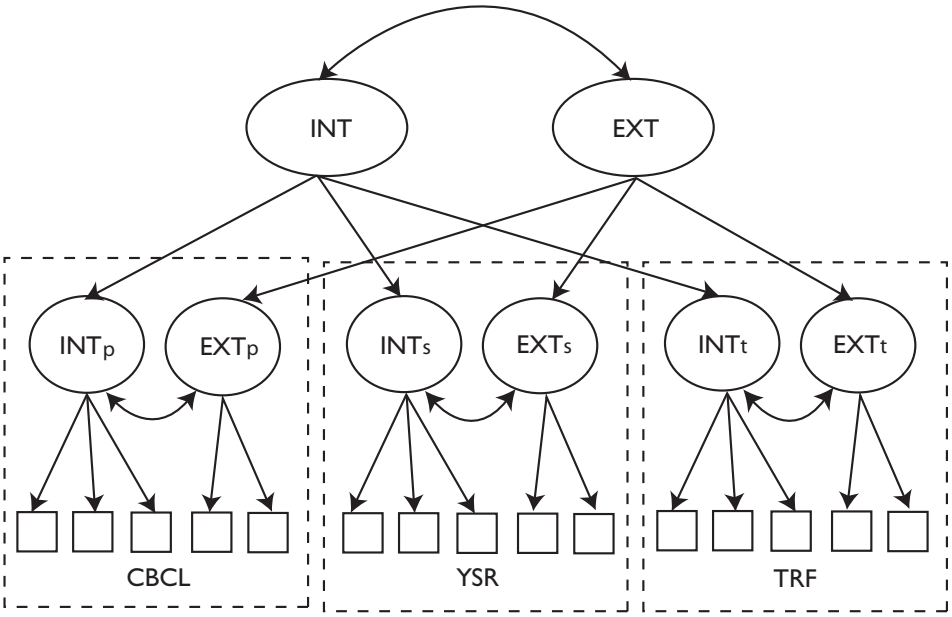

Figure 3. Multiple Indicator Correlated traits – Correlated Uniqueness model with three informants.

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.

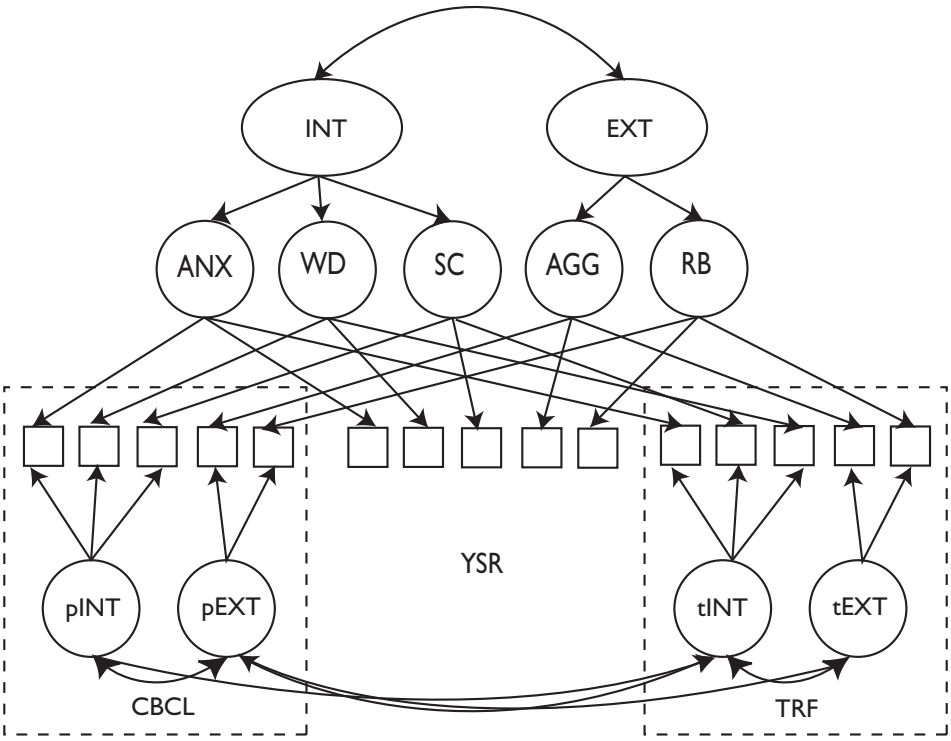

Figure 4. Correlated traits – Correlated Methods - 1 model with three informants. In this case the self-report is used as the reference method.
Note: INT = Internalizing; *EXT* = Externalizing; *INTp / EXTp* = Parent specific factors; *INTt / EXTt* = Teacher specific factors; *ANX* = Anxious-Depressed; *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.

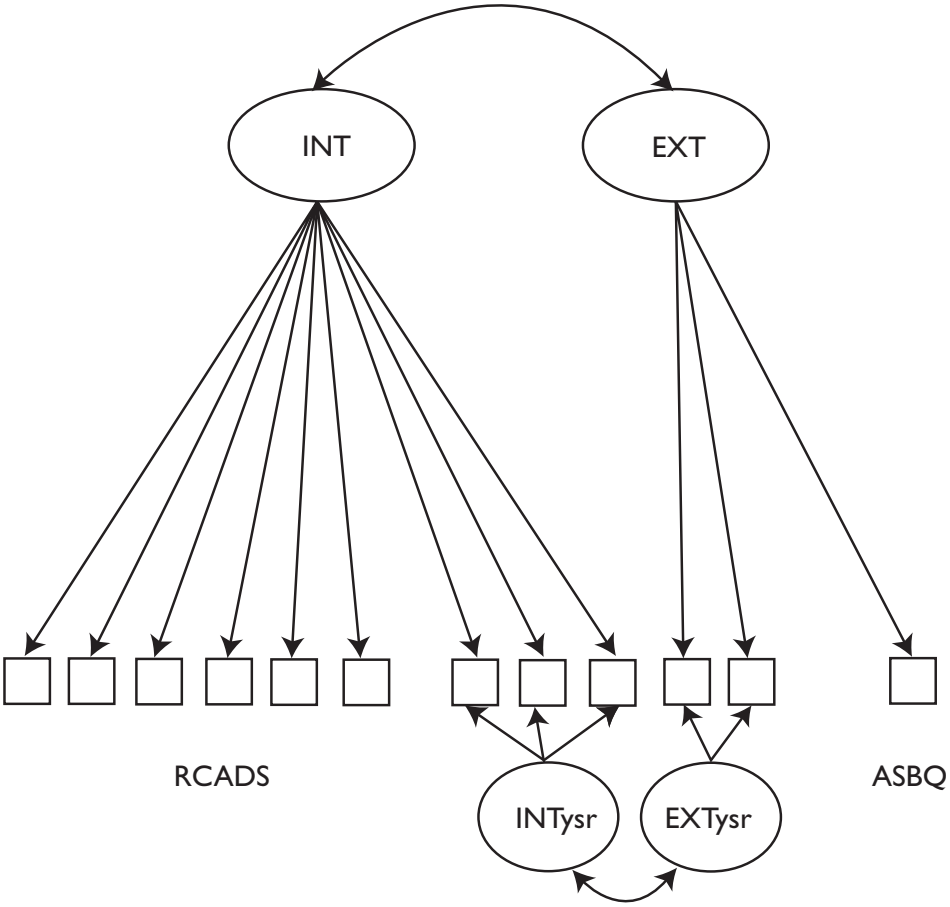

Figure 5. Correlated traits – Correlated Methods - 1 model with multiple instruments. Covariance specific to the Youth Self Report subscales is modelled as a specific method-factor.

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.