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).

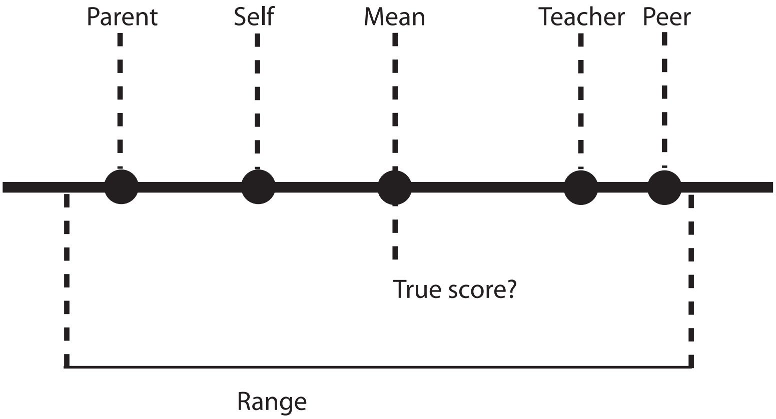

Figure 1. Illustration of the assumption that multiple informant-reports measure the same (dimensional) attribute.

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.

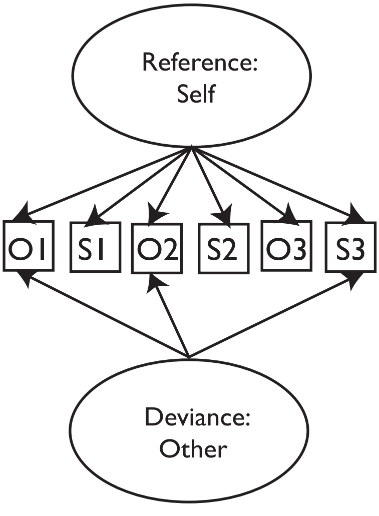

Figure 2. The CT-C(M-1) model with two informants and self-report as reference-method.

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.