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)

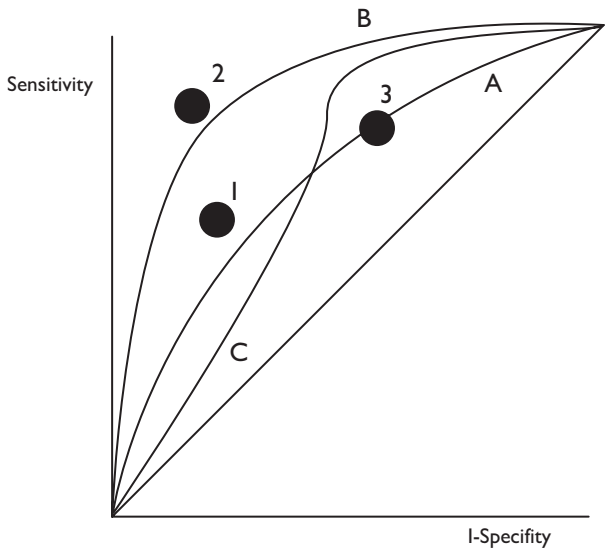

Figure 1. Receiver Operating Characteristics (ROC)-plane with some examples of ROC-curves and points.

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.