

DSM categories and dimensions in clinical and research contexts

HELENA CHMURA KRAEMER

Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

Abstract

An enhancement to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) is currently under consideration, one that would enhance both the reliability and validity of the Diagnostic and Statistical Manual (DSM) diagnoses: the addition of a dimensional adjunct to each of the traditional categorical diagnoses of the DSM. We first review the history and context of this proposal and define the concepts on which this dimensional proposal is based. The advantages of dimensional measures over categorical measures have long been known, but we here illustrate what is known with a theoretical and a practical demonstration of the potential effects of this addition. Possible objections to the proposal are discussed, concluding with some general criteria for implementing this proposal. Copyright © 2007 John Wiley & Sons, Ltd.

Key words: DSM, categorical, dimensional, diagnosis, power, precision

The Diagnostic and Statistical Manuals (DSMs) have long provided standard diagnostic guidelines for both the clinical and research use in psychiatry, although not without substantial criticisms. Some of the criticisms are certainly warranted, since the DSM continues a work in progress rather than a finished instrument. Some, however, arise because of semantic problems of which those seeking to develop DSM-V in the years to come must be aware.

The DSM concerns ‘diagnoses,’ not ‘disorders.’ A ‘disorder’ is something wrong in a patient that is of clinical relevance, a disease, a malfunction, an injury, an abnormality, etc., something problematic for the patient for which s/he would likely seek clinical attention, and for which clinicians might provide effective treatment. DSM focus and concern has always been on ‘diagnoses,’ that is, a clinical expert’s opinion as to whether some disorder is present in a particular patient. A diagnosis is to a disorder as a sample to the population it represents, or a measure to the construct it is meant to assess; that is, a representation or an indication, not the goal itself. When users forget the possibility of sampling error inherent in a sample, or the

measurement errors in a measure, results can mislead. In the same way, to date no diagnosis is perfectly reliable and valid for its disorder. Ignoring that fact can result in misleading conclusions.

The purpose of any diagnostic system, such as the DSM, is not to say what is ‘normal’ or ‘abnormal,’ nor what is or is not ‘acceptable’ in any society (Caplan, 1995); nor is it an effort to ‘medicalize’ society’s problems nor to channel clients to psychiatrists rather than to clinical psychologists, sociologists or other mental health providers (Kirk and Kutichins, 1992). DSM does not concern ‘insanity,’ a legal rather than a medical term, and assuredly does not concern who is ‘crazy’ or ‘mad,’ terms that are layman pejorative terms, not necessarily related to mental health disorders. Such terms continue to stigmatize those with mental health problems and are a major factor in the less than adequate care that those with mental health problems continue to receive. Yet many of the criticisms of DSM use exactly those terms, e.g. ‘They Say You’re Crazy: How the World’s Most Powerful Psychiatrists Decide Who’s Normal (The Inside Story of the DSM)’ (Caplan, 1995).

The word 'diagnostic' in DSM is clearly descriptive of its purpose to provide the best guidance currently available to identify those with a disorder. More puzzling is the word 'statistical' in DSM. Diagnostic and Statistical Manual of Mental Disorders, First Edition (DSM-I) and Second Edition (DSM-II) were proposed for purposes related primarily to counting cases: How many of those in institutions were in this general category rather than another? Is the number of those in a certain category increasing or decreasing over time? For such purposes, then and now, categorical diagnoses were necessary.

However, starting with DSM-III, it has been recognized that DSM diagnoses serve many other types of clinical and clinical research purposes as well. In the clinic, they serve purposes related to prevention, early identification, management, assessment of improvement, for individual patients. In clinical research, the diagnoses play important roles in seeking to understand etiology and course, to identify effective and cost-effective treatments, i.e., to provide the evidence for evidence-based clinical decision-making.

The DSM is not designed or intended to further basic science (the development of scientific theories, research on tissues or animal models), whether that basic science is medical, psychological, or sociological. It is not designed, for example, to further current medical, psychological, or sociological theories about cognition, personality, or functioning, although clearly all three are central to DSM diagnoses. If and when such theories lead to evidence that would advance understanding of the etiology, identification, treatment, course, or prognosis, then those results should and would influence DSM. Thus basic science should be expected ultimately to drive DSM, not the other way around.

With such issues in mind, clearly the word 'statistical' in DSM now takes on greater meaning, for a goal of DSM is to facilitate drawing correct statistical inferences from what is observed. In the clinic, that would mean correct inferences about choice of treatment, monitoring treatment response, maintaining health. In clinical research that would mean guidance on issues related to measurement, design, analysis, and clinical interpretation of research results, in epidemiological studies, in medical test evaluation, and in randomized clinical trials.

Obviously, while DSM relates to diagnosis and not to disorder, the goals of the DSM can be approached

only if there were close relationships between diagnoses and disorders. Such relationships are generally described by the reliability and validity of the diagnosis applied to a given disorder. Technically, the reliability of a diagnosis is the percentage of the person-to-person variability in a given population that relates to the variance of the 'true' values of the diagnosis (Lord and Novick, 1968). Less technically, it relates to the extent to which a second independent diagnostic opinion about a patient agrees with the first, and is best measured by the correlation coefficient between independent test-retest diagnoses for a sample of subjects from that population.

Validity, however, is the percentage of the person-to-person variability of the diagnosis in a given population that relates to the variance of the disease for which the diagnosis is meant, and is consequently always lower than the reliability of a diagnosis (Lord and Novick, 1968). To date, the DSMs have focused solely on face or clinical validity, the assertion that the diagnosis corresponds to clinicians' subjective views of a disorder. This is a weak but necessary form of validity achieved by requiring consensus among clinicians expert in that disorder, and such consensus has to date been the primary basis of DSM modifications. Ideally the validity of a diagnosis represents the correlation between the diagnosis and a 'gold standard' determination of the disorder. For example, one common form of validity is expressed by the sensitivity and specificity of a categorical diagnosis relative to its corresponding disorder, where sensitivity is the probability that a person who has the disorder is diagnosed positive, and specificity is the probability that a person who does not have the disorder is diagnosed negative. However, in absence of 'gold standard' diagnosis to serve as a criterion against which the diagnosis is tested, this is not yet possible for DSM, nor is it possible for most medical diagnoses.

In absence of a 'gold standard,' establishing validity means challenging the validity using a variety of external criteria. The more such challenges a DSM diagnosis can withstand, the more likely it is to be valid. For example, if those with a certain diagnosis have a subsequent course similar to each other and quite different from those without that diagnosis, this is a type of predictive validity. If those with a certain diagnosis have a different genetic profile, or are exposed to different environmental conditions, have a different characteristic brain structure and/or functioning, or respond

differently to treatment, than those without, each such demonstration provides increasing support for the validity of the diagnosis. If those with a diagnosis respond to certain interventions and not to others, that also is support for the validity of the diagnosis (Robins and Barrett, 1989; Robins and Guze, 1970).

Since DSM-III, major emphasis has been on test-retest reliability for three major reasons. First, reliability is easily assessed in practice, while, in absence of a 'gold standard' determining the presence or absence of the relevant disorder, establishing validity is a long and difficult process. Second, since the reliability of a diagnosis sets the ceiling for its validity, validity requires adequate reliability. Kirk and Kutichins (1992) criticized DSM because of this focus, suggesting that the development of the kappa as a measure of reliability had somehow derailed the diagnostic development process by focusing interest solely on reliability. This ignored the fact that another form of kappa serves as a primary measure of validity (Kraemer et al., 2002). Nevertheless their criticism of the exclusive focus to date on reliability is valid and motivates increased efforts to challenge the validity of the diagnoses in future DSM development. Finally, it is well known that unreliability attenuates effect sizes and decreases the power of statistical tests, both of which compromise the ability of research to provide the evidence necessary to validating the diagnosis or to guiding modifications to the DSM (Kraemer and Thiemann, 1987, 1989; Kraemer, 1991).

The process of DSM development is analogous to a spiral (Kupfer and Thase, 1989): A DSM version is proposed, which leads to clinical use and research applications that generate information on the reliability and validity of that diagnosis, which information is then used as a basis of the formulation of the next DSM version, and the process goes around again and again. Each successive iteration is expected to move closer to the true disorder. Where DSM-I and DSM-II focused on clinical validity, and DSM-III emphasized reliability, DSM-IV began to emphasize an evidence-based approach to diagnosis. DSM-V would be expected to do all this and more.

We propose one enhancement to DSM for DSM-V, one there is reason to believe would enhance both the reliability and validity of DSM diagnoses: the addition of a dimensional adjunct to each of the traditional categorical diagnoses of the DSM. Including a corresponding dimensional scale along with a categorical

diagnosis in DSM-V has nothing to do with the nature of disorders, and everything to do with the quality of a diagnosis for that disorder and for the clinical and research needs that such a diagnosis might serve. DSM's exclusive focus so far on categorical diagnosis is a historical fact stemming from the goals of the earliest DSMs. As the goals of DSM have broadened, the need for a corresponding dimensional approach has become more urgent.

In what follows we will first define the concepts on which this dimensional proposal is based. Then we will show a theoretical and a practical demonstration of the potential effects of this proposal, discuss possible objections, and consider right and wrong ways to implement the proposal.

Categorical and dimensional diagnosis

A categorical diagnosis (at least in the way that term is used in DSM) has only two values: The patient is either positive (thought to have the disorder) or negative (thought not to have the disorder). Generally a categorical measure is one with two or more discrete non-ordered responses, and, technically, DSM uses binary diagnoses, but to avoid confusion we will continue to use the term most often used: categorical diagnosis.

A dimensional scale in contrast has three or more ordered values. Thus a three-point scale (e.g. 0, none; 1, some; or 2, severe symptoms) is dimensional (although little better than a categorical diagnosis), as is a 4, 5, 6, 7, . . . point scale (e.g. the Hamilton Depression Scale), a discrete score (e.g. number of drinks per week), or a continuum (e.g. Body Mass Index, duration of symptoms). Multivariate diagnoses are included here as well (e.g. the combination of age of onset, total duration, and severity of symptoms). Again 'dimensional' is perhaps here a misnomer, for what is proposed would better be described as 'ordinal' (univariate or multivariate), but we will continue to use the usual term 'dimensional.'

What is being proposed for DSM-V is not to substitute dimensional scales for categorical diagnoses, but to add a dimensional option to the usual categorical diagnoses for DSM-V. While it has often been opined that clinicians want categorical diagnoses and researchers want dimensional ones, that is not precisely true (Kraemer and O'Hara, 2004). To assess whether a patient is responding to treatment, clinicians value the information provided by dimensional approaches. To

decide whether a patient is eligible for a research study, researchers typically rely on a categorical diagnosis. Thus, if a dimensional option is added, those clinicians and clinical researchers who need or prefer to use the categorical diagnoses in certain contexts would continue to do so. Where corresponding dimensional scales enhance clinical or research decision-making, those dimensions would become available in DSM-V.

When is a dimensional diagnosis unneeded or impossible?

The brief answer: virtually never. The only situation in which a dimensional adjunct would not add quality to the categorical diagnosis for a disorder is when there is no meaningful clinical variation among those who are diagnosed positive on the categorical diagnosis, and no meaningful clinical variation among those who are diagnosed negative on the categorical diagnosis.

Consider this: In the past, cancer treatments were often compared using the success rates. If one survived 5 years past diagnosis, one was a success; otherwise a failure. What this categorical definition did was to equate surviving 5 years plus 1 day to surviving 50 years post diagnosis, and to equate surviving 5 days post diagnosis to surviving 5 years minus 1 day. However, those surviving 5 years plus 1 day were considered completely different from those surviving 5 years minus 1 day, as different as those surviving 50 years were from those surviving 5 days. To both patient and clinician that makes little practical sense. In more recent years, survival methods are used instead of this categorical response, thus substituting a dimensional measure (time) for a categorical measure.

That illustrates the problem with every categorical DSM diagnosis. Among those who have the diagnosis, there is variation in precursors: genotype, environmental exposures, age of onset, pre-morbid physiological, psychological, behavioral and emotional characteristics; in concomitants: specific symptomatology, severity, duration of episodes and remissions, response to treatments; and in consequences: disability, impairment, diminished quality of life, shortened lifespan. Those with the diagnosis differ from each other in many respects, but which of the many ways in which those with the disorder differ are clinically significant is the question important to DSM-V. Similarly, those without the diagnosis differ from each other in the same precursors and many concomitants which may indicate resistance or resilience to the disorder of

interest. Which of these are clinically significant only adds to the puzzle.

Thus every DSM categorical diagnosis would be enhanced with a dimensional adjunct, and the challenge to choose the appropriate dimensional diagnosis for each disorder is to identify what are the most clinically important sources of heterogeneity among those who have the categorical diagnosis, and among those who do not. Is it the heterogeneity in severity of symptoms, or is it the impairment induced? Is it the consistency with which symptoms are expressed, or the duration of symptoms? Is it the inability to inhibit symptom expression, lack of control? All of these and more are possibilities that the experts in each disorder must consider.

But perhaps for some DSM categorical diagnoses, adding a dimensional diagnosis might be desirable, but not possible? Current DSM categorical diagnoses are already frequently based on dichotomizing dimensional diagnoses. Consequently, it is difficult to argue that developing dimensional diagnoses cannot be done. Every categorical diagnosis can be made dimensional by using symptom counts, symptom duration, symptom severity, degree of impairment, certainty of diagnosis, consensus of multiple diagnoses, and many more such strategies even without deviating from the contents of current DSM categorical diagnoses. Thus the issue is not whether a dimensional diagnosis can be added to each categorical diagnosis, it is merely how best to do that for each.

Two illustrations of impact

A thought experiment

There has long been an extensive literature documenting the advantages of dimensional over categorical measures in research, but the potential impact still seems not well understood. To give a dramatic example, suppose there were a mental health disorder that was caused by a single gene: i.e. if we could get an error-free measure of the expression of that gene, G^* , and an error-free dimensional diagnosis of the disorder, D^* , the correlation between G^* and D^* would be perfect. The problem is, of course, we cannot get error-free measures of either. Instead suppose G and D are obtainable dimensional measures of G^* and D^* with a validity of, say, 0.6. Then the correlation between G and D is not 1.0, but 0.6. Whereas, it would take a sample size of no more than 10 to detect that perfect correlation between

G^* and D^* , to achieve 80% power with a 5% one-tailed test to detect positive association between G and D would require only $N = 13$.

However, suppose now we dichotomized G , to represent the presence/absence of the single gene that determined G^* , and D for a categorical diagnosis. In Figure 1 is shown the total sample size necessary to have credibility and adequate power to detect a positive association when the proportion 'with the gene' is Q and the proportion with a positive diagnosis is P . As can be seen, the optimal dichotomization is when both G and D are dichotomized at their medians, in which case the sample size necessary for 80% power is approximately three times as great as that using dimensional G and D . When the dichotomization is less than optimal resulting in small P and Q , as so often happens with diagnoses, the sample size necessary for 80% power may be five or even 10 times as large as that necessary using the dimensional G and D . In this illustration, we start with perfect correlation and with unavoidable unreliability of measurement coupled with categorical measures, end with a relation difficult to document as non-random. Moreover since unreliability and dichotomization both attenuate effect size, the actually perfect correlation between G^* and D^* might in the end be characterized as 'small' or, at best, 'moderate' and often reported as not statistically significant association. So easy is it to cause even perfect association between a gene and a disorder to disappear.

There is a considerable methodological literature documenting the losses associated with using a categorical variable when a dimensional variable is available (Kraemer and Thiemann, 1987; Cohen, 1983; Donner and Eliasziw, 1994). It is a rare circumstance, likely only in absence of clinically relevant heterogeneity among those with and without the dichotomous diagnosis, when for purposes of inference in the clinic or in research, there is not a loss of precision and power in using the dichotomous diagnosis.

A practical example

But, some would protest, that could not happen in reality. Theoretical arguments are often not quite as convincing as a real example. Consider then the following: Agras et al. (personal communication, manuscript in preparation) conducted a four site randomized clinical trial to test the relative effectiveness of Cognitive Behavior Therapy (CBT) against self-help for eating disorders, and, more specifically to test whether a risk classification at entry to the study moderated (14) the effect of CBT versus self-help, as suggested from exploratory studies in earlier randomized clinical trials (Kraemer et al., 2006). In this study, there were 285 subjects distributed over the four sites. The outcome measure, as is typical in this field, was a categorical one, where success was defined as Binge Frequency and Purge Frequency below specified cut-points, with little or no empirical evidence documenting that

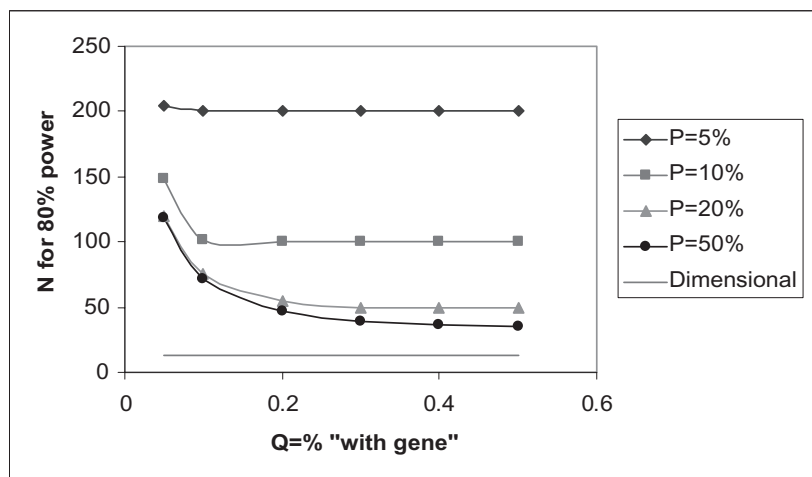


Figure 1. The sample size necessary to achieve 80% power with a 5% one-tailed test to detect a positive correlation between 'the gene' and a categorical diagnosis where the percentage with the diagnosis is P , and the percentage with 'the gene' is Q . Also shown is the sample size necessary if using a dimensional diagnosis.

the cut-points were optimal. No significant treatment effect was found, nor was any significance found for the proposed moderator of treatment. Since non-significant effects cannot be interpreted as support for the null hypothesis, the result of this study was a 'hung jury'.

Suppose instead, we were here to use a dimensional measure, the weighted sum of the binge and purge frequencies maximally correlated with the usual categorical outcome (developed using Logistic Regression with the categorical outcome as the dependent measure with binge and purge frequencies as the independent measures). If the analysis were done using the dimensional measure (same subjects, same treatments, same design), there is a statistically significant moderator effect of risk strata ($p = 0.023$). Indeed in the low risk stratum, CBT was better than self-help with $NNT = 9$, but in the high risk stratum the result was in the opposite direction: self-help was better than CBT again with $NNT = 9$. Such a finding is important since it suggests different choices of treatment for those in the high and low risk strata.

What was seen in this study is not unusual. In many of the results of randomized clinical trials or of risk studies that use categorical measures, a report of statistical non-significance may be partially or wholly due to the lack of power to detect effects due to use of categorical measures, particularly when the cutoff defining the categorical measures is set by intuition rather than optimally based on empirical evidence. Conversely, the reason sample sizes must be as large as they are (thus taking more time and requiring greater funding investment) may be related to the pervasive use of categorical diagnoses.

What if this proposal were accepted: objections

The proposal to include dimensional diagnoses into DSM-V is not without objections, well represented by those of Michael First (2005). First (2005) points out the lack of data about clinical utility and user acceptability, and clearly both are important. However, until a dimensional diagnosis is included in a version of DSM, there can be no documentation of either the clinical utility or user acceptability. First (2005) also suggests that inclusion of a dimensional diagnosis will complicate medical record keeping and create administrative and clinical barriers between mental disorders and medical conditions. However, most medical categorical diagnoses already include a dimensional diagnosis as well. When the physician gives a diagnosis of

hypertension, the systolic and diastolic blood pressure are recorded; a diagnosis of breast cancer is extended by noting the Stage and/or the Karnofsky score; a diagnosis of diabetes, by noting the fasting blood sugar level, etc. First (2005) suggests that inclusion of a dimensional diagnosis in DSM-V would require a massive re-treading effort. Any change in a diagnostic system as established as the DSM is always disruptive. However, mental health providers have long used individually chosen dimensional measures (e.g. the Hamilton score for depression). What inclusion of a dimensional diagnosis in DSM-V would accomplish would be to coordinate the use of dimensional diagnoses across clinicians with the benefits obtained similar to those obtained when the use of categorical diagnoses was coordinated in the earliest DSMs. Similarly First (2005) suggests that adding a dimensional component might disrupt research efforts, for example, making it impossible to pool studies using DSM-IV and DSM-V. Unfortunately this is true, but pertains to any change in the categorical diagnoses as well. Any change in the criteria raises issues as to whether research done with one version of DSM can be validly pooled with another version of DSM. Since a dimensional diagnosis in DSM-V would not displace the categorical diagnosis, introduction of a dimensional diagnosis would have less impact in this regard than does any modification of the categorical diagnosis, and there will undoubtedly be such modification. Similarly First's (2005) concern that inclusion of a dimensional component would complicate clinicians' efforts to integrate prior clinical research using DSM categories into clinical practice, applies more to the modification of the categorical diagnosis than it would to an added dimensional diagnosis. Research using the dimensional diagnosis would only begin after DSM-V.

What if this proposal were accepted? Criteria and caveats

To achieve what is hoped from addition of a dimensional diagnosis into DSM-V, there are certain criteria that would have to be met, and certain approaches may be less than fruitful.

- It is crucial that a DSM-V dimensional diagnosis correspond well with its categorical diagnosis. Those with a positive categorical diagnosis should have dimensional diagnoses much higher than those with a negative categorical diagnosis. Conversely, if one stratifies the population on the dimensional

diagnosis, the probability of a positive categorical diagnosis should consistently increase as the dimensional diagnosis increases. If this were not so, research using the categorical and dimensional diagnoses ostensibly for a single disorder would actually deal with two different disorders. It would be difficult, if not impossible, to reconcile the conclusions from studies. With a close positive correlation between the categorical and dimensional diagnoses, the major source of discrepancy between studies done using the two will lie in the greater power and precision of studies using the dimensional diagnosis.

- DSM is meant for the use of clinicians and thus must be transparent to clinicians. Complex algorithms for computation of a dimensional diagnosis will not be useful to clinicians, and research based on such diagnoses will not be easily interpretable by clinicians. Consider the Framingham Index, a dimensional measure indicating risk for cardiac events (Truett et al., 1967). This Index requires, for example, that the clinician assess the patient's age, gender, smoking status, whether or not the patient is on medication for high blood pressure, total cholesterol, HDL cholesterol, systolic blood pressure. From this information a weighted score is computed from which the probability of a future cardiac event can be estimated. Score sheets are available and calculators appear on the Web, but the calculation of such a risk score has not become a routine procedure for clinicians. However, a simplified version in which each of the indicators is dichotomized and given a point count indicating the importance of each indicator, with the total point count convertible to a risk level, is also available. This simplified version of a dimensional score may not be as accurate a predictor of cardiac risk, but because the clinician can see exactly what goes into the score and needs no complex computation procedure, it is more likely to be used in practice. In the same way, every effort should be made to make the DSM-V dimensional diagnosis clinically useful, if not necessarily mathematically elegant. Consequently, there are certain paths to the development of dimensional diagnosis that must be viewed with concern. While completely subjective and non-evidence based decision-making is likely to subvert the process, complex mathematical models, based on assumptions that may or may not be true, may

generate dimensional diagnoses not only non-transparent to clinicians, but very possibly non-valid. At the other extreme, complex mathematical methods (latent variable modeling, item response theory, etc.) are valuable tools to be used to help generate possible approaches to dimensional diagnoses, but if what results does not correspond to clinical knowledge and insight, and is not verified for test–retest reliability and validity against external criteria, these results may also subvert efforts.

- Finally, the dimensional diagnosis, as is also true of the categorical diagnosis, must show good test–retest reliability, have clinical validity, and withstand more demanding challenges to validity. Consequently, the dimensional diagnosis, as is also true of the categorical diagnosis, must be evidence-based, reflecting what has been learned about diagnosis in the years since the introduction of the DSM-IV. The emphasis in the evidence must be on effect sizes, which indicate clinical significance, and not on statistical significance (p -values) for all the reasons made ever clearer in recent years of the limitations of statistical hypothesis testing when the focus is on p -values rather than effect sizes (Borenstein, 1997; Cohen, 1995; Dar et al., 1994; Hunter, 1997; Jacobson and Truax, 1991; Kraemer, 1993; Kraemer et al., 1999; Schmidt, 1996; Shrout, 1997; Thompson, 1999; Wilkinson, 1999).

Discussion and conclusion

The discussion about adding a dimensional component to the DSM has been ongoing for many years. In the earliest stages of development of each version of DSM since DSM-III, the ideal has surfaced, and then again submerged. The time to include a dimensional diagnosis is now, with DSM-V, not only because the arguments for doing so are so strong, but also to begin to prepare for the possible inclusion of genetic, imaging, biochemical, or other signals into future diagnostic systems.

The process of adding a dimensional component to the DSM at this time can be made as simple or as complicated as is appropriate given the state of knowledge about each disorder. Personality disorders, for example, have long focused on dimensional diagnosis and are likely more than ready to move in this direction. Other diagnoses have given the idea and process less thought, and may be at an earlier phase of the process. The first version of a dimensional diagnosis, like the earliest versions of categorical diagnoses, is

unlikely to be anywhere near perfect. However, as dimensional diagnosis is proposed, used and evaluated, it too, like categorical diagnosis, will improve with time and the accrual of evidence based on those diagnoses.

Nevertheless, the consequences of including a dimensional component in DSM-V may revolutionize psychiatric research and clinical decision-making, and bring progress in dealing with mental disorders more in line with progress in dealing with other medical conditions.

References

- Borenstein M. Hypothesis testing and effect size estimation in clinical trials. *Annal Allergy, Asthma, and Immunology* 1997; 78: 5–16.
- Caplan PJ. They say you're crazy: how the world's most powerful psychiatrists decide who's normal. Cambridge, MA: Da Capo Press, 1995.
- Cohen J. The cost of dichotomization. *Appl Psychological Measurement* 1983; 7(3): 249–53.
- Cohen J. The Earth is round ($p < 0.05$). *Am Psychologist* 1995; 49: 997–1003.
- Dar R, Serlin RC, Omer H. Misuse of statistical tests in three decades of psychotherapy research. *J Consulting Clin Res* 1994; 62: 75–82.
- Donner A, Eliasziw M. Statistical implications of the choice between a dichotomous or continuous trait in studies of interobserver agreement. *Biometrics* 1994; 50: 550–5.
- First MB. Clinical utility: a prerequisite for the adoption of a dimensional approach in DSM. *J Abnorm Psychol* 2005; 114(4): 560–4.
- Hunter JE. Needed: a ban on the significance test. *Psychological Sci* 1997; 8(1): 3–7.
- Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consulting Clin Psychol* 1991; 59(1): 12–19.
- Kirk SA, Kutchins H. *The Selling of the DSM: The Rhetoric of Science in Psychiatry*. New York: Aldine De Gruyter; 1992.
- Kraemer HC. Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. *Psychoneuroendocrinology* 1993; 17: 527–36.
- Kraemer HC. To increase power without increasing sample size. *Psychopharmacology Bulletin, Special Feature: ACNP Proceedings* 1991; 27(3): 217–24.
- Kraemer HC, O'Hara R. Categorical versus dimensional approaches to diagnosis: methodological challenges. *J Psychiatric Res* 2004; 38(1): 17–25.
- Kraemer HC, Thieman S. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications; 1987.
- Kraemer HC, Thieman SA. A strategy to use 'soft' data effectively in randomized clinical trials. *J Consulting Clin Psychol* 1989; 57: 148–54.
- Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. *J Am Med Assoc* 2006; 296(10): 1–4.
- Kraemer HC, Periyakoil VS, Noda A. Tutorial in biostatistics: Kappa coefficients in medical research. *Statistics in Medicine* 2002; 21: 2109–29.
- Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. *Psychological Methods* 1999; 4(3): 257–71.
- Kupfer DJ, Thase ME. Laboratory studies and validity of psychiatric diagnosis: has there been progress? In Robbins LN, Barrett JE (eds) *The Validity of Psychiatric Diagnosis*. New York: Raven Press, 1989, pp. 177–201.
- Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company, 1968.
- Robins LN, Barrett JE (eds). *The Validity of Psychiatric Diagnosis*. New York: Raven Press, 1989.
- Robins E, Guze SB. Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. *Am J Psychiatry* 1970; 126: 983–87.
- Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods* 1996; 1(2): 115–29.
- Shrout PE. Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Sci* 1997; 8(1): 1–2.
- Thompson B. Journal editorial policies regarding statistical significance tests: heat is to fire as p is to importance. *Educational Psychol Rev* 1999; 11: 157–69.
- Truett J, Cornfield J, Kannel WA. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chronic Dis* 1967; 20: 511–24.
- Wilkinson L. The Task Force on statistical inference. *Statistical methods in psychology journals: guidelines and explanations*. *Am Psychologist* 1999; 54: 594–604.

Correspondence: Helena Chmura Kraemer, Department of Psychiatry and Behavioral Sciences, Stanford University, 401 Quarry Road, MC 5717, Stanford, CA, 94301, USA.

Phone: 650 328-7564

Email: hckhome@pacbell.net