The Categorical versus Dimensional Assessment Controversy in the Sociology of Mental Illness*

RONALD C. KESSLER

Harvard Medical School


This paper reviews the sociological controversy over using categorical versus dimensional assessments in the study of psychological distress. The preference of sociologists for dimensional assessments is traced to two assumptions: (1) that the associations of predictors with psychological distress syndromes are most accurately operationalized by using dimensional measures; and (2) that no true discrete mental illnesses can reasonably be inferred to exist that would justify the creation of dichotomous measures. Methods are described in this paper to test both assumptions. An argument is made that the first assumption is the critical one and that dimensional analysis is useful only when it can be demonstrated that the associations of predictors with dimensional scores are constant throughout the relevant dimensional severity range. The paper closes with an argument for the tandem use of categorical and dimensional assessments in future sociological research.

The controversy surrounding the use of categorical assessments (yes/no decisions regarding whether a person does or does not have a mental illness) versus dimensional assessments (each person receiving a score on a continuous scale of psychological distress with no cut-point to designate a threshold between those with and without a presumed illness) to create outcomes in sociological studies of mental illness can be traced back nearly half a century to the early community psychiatric studies of mental illness that were carried out shortly after World War II (Leighton et al. 1963; Srole et al. 1962). Many sociologists came down on the side of dimensional assessment at that time, and many sociologists have continued to use dimensional measures of distress as the main outcomes in their community surveys of mental illness ever since. The two main reasons advanced for preferring dimensional to categorical measures are that (1) the relationships between predictors and syndromes of psychological distress are more accurately captured in statistical models that specify dimensional rather than categorical representations of distress and that (2) there is no evidence for the existence of true discrete mental illnesses that account for the patterns observed among symptoms in dimensional assessments. Mainstream psychiatric epidemiologists (some of whom were trained as sociologists), in comparison, moved away from an initial use of dimensional assessments to categorical assessments, and then to a more recent use of both types of assessments in tandem. I argue in this paper that the tandem approach makes more sense than an exclusive use of either dimensional or categorical assessments. I also argue that the decision to use either categorical, dimensional, or both types of assessment should be made independent of whether the researcher believes that there is a true discrete illness that accounts for the symptoms under investigation.

* Preparation of this report was supported by U.S. Public Health Service Grants U01 MH60220 and R01 DA12058 and by the Pfizer Foundation. Empirical results reported here are based on the National Comorbidity Survey (NCS) and the NCS Replication (NCS-R). More details on the NCS, the NCS-R, and the screening scales described in the text based on these surveys can be obtained from the web at www.hcp.med.harvard.edu/ncs. Address comments to Ronald C. Kessler at the Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Suite 215, Boston, MA 02115. Voice: 617-432-3587, Fax: 617-432-3588, Email: kessler@hcp.med.harvard.edu.

171
BACKGROUND

The earliest dimensional scales of nonspecific psychological distress used in community epidemiological surveys were the 22-item Langner (1962) Scale used in the Midtown Manhattan Study (Srole et al. 1962) and the 20-item Health Opinion Survey (Macmillan 1957) used in the Stirling County Study (Leighton et al. 1963). Both the Langner and Health Opinion Survey scales were based on a more detailed screening scale, the Neuropsychiatric Screening Adjunct (Star 1950), developed for Selective Service screening during World War II. These early scales were all designed to be first-stage screening tools used to target respondents with broadly defined emotional problems for more in-depth clinical assessment. This comparison between screening scores and clinical evaluations of mental illness led to the establishment of optimal cut-points on the screening scales for differentiating “cases” and “noncases.” These validated cut-points were then used in later surveys to report prevalence rates (i.e., percentages of respondents with mental illness), without clinical follow-up to transform dimensional screening scale scores into dichotomous case definitions. However, controversies arose regarding the appropriate cutpoints on these scales for case thresholds in community surveys (Seiler 1973). These controversies were fueled by the absence of a consistent way of defining clinical cases in the validation phase of the research, which led to inconsistencies in prevalence estimates across surveys (Dohrenwend and Dohrenwend 1974). In addition, differences in prevalence across community samples and clinical samples led to differences in the positive predictive values of screening scale scores (the proportion of respondents with a given screening scale score who would be confirmed as a “case” by a clinical interviewer) across samples. Failure to account for differences in positive predictive values introduced further imprecision into community survey prevalence estimates. In the face of the confusion caused by these problems, and in the presence of evidence that the symptoms included in these screening scales have a strong unidimensional component (Dohrenwend et al. 1980), these screening scales came to be reported only in dimensional form (e.g., means, plots of distributions) in later surveys (e.g., Myers, Lindenthal, and Pepper 1975).

The cut-point debate focused initially on a narrow clinical question: “What is the correct cut-point to define clinical significance of psychiatric symptoms?” However, the labeling theory debate in sociology (Gove and Howell 1974; Scheff 1974), which was going on at roughly the same time as the cut-point debate, raised a deeper question: “Does it make sense to believe that there is any true illness that provides a principled basis for deciding on a cut-point?” The recognition that the diagnostic criteria for mental disorders stipulated in the American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders (DSM) often had as much to do with votes by committees as scientific evidence (Kirk and Hutchins 1992) added a conceptual rationale to the predisposition of sociologists to favor the use of dimensional rather than categorical assessments. The situation has not changed much in the intervening three decades. Despite a dramatic expansion in our knowledge of the neurobiology of psychiatric syndromes (Charney 1999), psychiatry continues to be unable to develop definitive biological tests for any mental disorder. As a result, the preference for dimensional over categorical assessments of these syndromes among sociologists continues to this day. There is also a strong movement within psychiatry for favoring dimensional measures over arbitrary categories based on similar reasoning in addition to an emphasis on clinical usefulness (Goldberg 2000).

Link and Dohrenwend (1980), in a detailed review of the content and interpretation of dimensional screening scales of nonspecific psychological distress, showed that these scales typically include questions about a heterogeneous set of cognitive, behavioral, emotional, and psychophysiological symptoms that are elevated among people with a wide range of different mental disorders. These authors also demonstrated that, despite this heterogeneous content, the vast majority of the symptoms in these scales have high factor loadings on a first principal factor (Link and Dohrenwend 1980): This finding suggests that all these screening scales tap a single broad underlying dimension of nonspecific psychological distress. It is not surprising, based on these findings, that studies of correlations among categorical measures of psychiatric diagnoses in psychiatric epidemiological sur-
veys find high rates of comorbidity (Kessler 1995; Robins, Locke, and Regier 1991).

Despite this evidence of a strong unidimensional core to psychological distress, there are meaningful secondary factors in psychometric studies of screening scale scores. Consistent with this finding, psychiatric epidemiological surveys find that a sizeable number of people have pure categorical mental disorders (i.e., one and only one disorder). Based on this kind of evidence, researchers who favor dimensional assessment have moved beyond an exclusive interest in measures of psychological distress to study distinct dimensional measures of anxiety, depression, and several of the other syndromes that have been differentiated in psychometric research on the structure of psychological distress (Derogatis 1983; Dohrenwend et al. 1980). However, few sociological studies examine categorical diagnostic transformations of these scales.

The situation is quite different in mainstream psychiatric epidemiology, which moved away from the use of screening scales of non-specific psychological distress in community surveys beginning in the early 1980s. This move occurred as a result of the influential Epidemiological Catchment Area (ECA) Study (Robins and Regier 1991). It is noteworthy that a number of sociologists played key roles in the ECA Study. The cut-point problem was resolved in the ECA study by taking advantage of the fact that the newly created DSM-III diagnostic system (American Psychiatric Association 1980) provided rules for operationalizing diagnostic criteria that were much clearer than in previous versions of the DSM. These rules were used to develop a fully structured interview schedule (i.e., an interview schedule that could be used by trained interviewers who are not clinicians), the Diagnostic Interview Schedule (Robins et al. 1981), to generate diagnoses according to the definitions and criteria of DSM-III. It was only due to the fact that the Diagnostic Interview Schedule was easy to administer and did not require clinical interviewers that the massive ECA study (over 20,000 respondents selected across five different sites interviewed at multiple points in time) was logistically and financially feasible. Importantly, the Diagnostic Interview Schedule was argued to have acceptable concordance with independent DSM-III clinical diagnoses based on subsequent blind ECA reinterviews carried out by psychiatrists (Helzer et al. 1985), although the empirical data comparing Diagnostic Interview Schedule and independent clinical diagnoses actually showed concordance to be rather poor (Anthony et al. 1985).

The ECA was an enormously influential study in psychiatry. A number of Diagnostic Interview Schedule surveys based on the ECA were subsequently carried out in countries throughout the world (e.g., Bland, Orn, and Newman 1988; Canino et al. 1987; Hwu, Yeh, and Chang 1989; Lee et al. 1990; Lépine et al. 1989; Wells et al. 1989; and Wittchen et al. 1992). Cross-national comparative studies of these surveys showed impressive consistencies in patterns and correlates of disorders (Cross-National Collaborative Group 1992). However, as most countries around the world use the World Health Organization’s (WHO) International Classification of Diseases rather than the DSM system to diagnose mental disorders, there was a need for a fully structured interview to generate diagnoses based on the definitions and criteria of the International Classification of Diseases. Recognizing this need, collaboration was established between the WHO and the developers of the Diagnostic Interview Schedule to create a broader interview that would generate both International Classification of Diseases and DSM diagnoses. The resulting instrument, the World Health Organization Composite International Diagnostic Interview (Robins et al. 1988), was subsequently used in a number of community epidemiological surveys around the world (World Health Organization International Consortium of Psychiatric Epidemiology 2000). Composite International Diagnostic Interview diagnoses have generally been shown in these surveys to have acceptable concordance with clinical diagnoses based on blind clinician reinterviews (Wittchen 1994).

It is important to realize that the impressive body of work that has occurred since the early 1980s in community epidemiological studies of categorical diagnoses depended fundamentally on the existence of a fully structured diagnostic interview like the Diagnostic Interview Schedule and that the creation of the Diagnostic Interview Schedule would not have been possible without the existence of the clear diagnostic classification rules that were first introduced in 1980 in the DSM-III. Agreement among clinicians and clinical researchers to adopt these rules led to a resolution of the
question raised by psychiatric epidemiologists in the early post World War II categorical-dimension assessment debate regarding the appropriate place to make clinical cut-points. However, it did not address the deeper question asked by sociologists regarding whether categorical distinctions make sense at all. The DSM-III categorical distinctions were made based on practical considerations having to do with presumed magnitudes of clinically significant distress and impairment combined with evidence regarding treatment response, rather than based on biological evidence justifying the diagnostic distinctions. The developers of the DSM-III were also sensitive to the fact that the system would be used for insurance billing purposes. This explains why so many "not otherwise specified" categories are included in the classification scheme. This creation of a document that met many needs led Jablensky (1999), a psychiatrist intimately involved in the classification process, to say that the DSM-III and later modification of the DSM and International Classification of Diseases systems created an agreement on nomenclature but not a logical system for classifying mental disorders.

Interestingly, even though the DSM-III system set the stage for the dominance of categorical classification in psychiatric epidemiological research over the next two decades, clinicians have continued to make use of dimensional scales, both to screen for mental illness in primary care (Goldberg 1972) and to assess symptom severity and treatment effectiveness in clinical studies (Rush et al. 1996). Similar to the screening scales developed shortly after World War II, the screening scales used in primary care are designed to be short, nonspecific measures that provide broad-gauged assessment of the presence of any mental disorder. More detailed clinical assessment is required among patients who screen positive to determine which disorders they have. The screening scales used to study symptom severity and treatment effectiveness, in comparison, are designed for use only among people known to have a specific disorder, rather than in the general population. This means that the questions in these clinical scales are much more specific to particular disorders than the questions included in scales designed for use in the general population. In addition, the thresholds of the questions in these clinical scales are a good deal higher than those found in the scales designed for use in the general population.

Many clinical decisions are based on the results of dimensional rather than categorical assessment, albeit in patient samples defined on the basis of prior categorical evaluations (Andrews forthcoming). For example, clinical experience shows that the usefulness of drug therapy in conjunction with behavioral therapy, rather than behavioral therapy alone, to treat phobias is related to the severity of phobic reactions assessed on dimensional scales such as the Marks Fear Questionnaire (Marks and Mathews 1979) and the Liebowitz Social Anxiety Scale (Liebowitz 1987). These scales are only administered after first determining, on the basis of an initial categorical clinical interview, that the patient does meet criteria for a phobia. Although it is possible, at least in principle, to carry out both types of assessment with a single dimensional scale that evaluates a clinical threshold and also evaluates symptom severity above that threshold, this is not done in practice because the types of information obtained in the two phases of assessment do not lend themselves to a consistent response format. In the categorical assessment, for example, it is usually quite important to ask about age of onset, course, family history, and range of trigger stimuli that lead to phobic fear and avoidance. In the symptom severity assessment, in comparison, the interviewer begins with the recognition that a phobia exists and asks the respondent to focus on the severity of their fear during exposure and the persistence of their efforts to avoid exposure to the feared stimuli. It is much more convenient to carry out these categorical and dimensional assessments separately rather than to devise a single dimensional assessment that can be administered to an unrestricted general population sample.

EVALUATING THE EXISTENCE OF DISCRETE ILLNESSES

As noted in the introduction, uncertainty regarding the existence of true discrete mental disorders has played a major part in the sociological preference for dimensional measures over categorical measures. Indeed, many sociologists who work with dimensional scales of anxiety, depression, and other psychological distress syndromes say that their main reason
for doing so in preference to categorical assessments of these same syndromes is that they do not believe that there is any evidence for the existence of a true dichotomous illness entity underlying the dimensional scores. It is consequently important to be aware of the fact that taxometric methods now exist that make it possible to test the hypothesis that dimensional symptom scale data are derived from an underlying discrete entity known as a taxon (Grove and Meehl 1993). Sociologists appear to be unaware of the existence of these methods.

It is also important to recognize that the existence of a taxon does not mean that people in the latent class are ill. It only means that a dimensional characterization of the indicators used to define the phenomenon under investigation will be inaccurate. External criteria are needed to determine whether given categories on discrete measures or cut-points on dimensional measures should be defined as indicating the existence of illness. These external analyses would be used to link either discrete or dimensional characterizations with measures of pain, suffering, role impairment, mortality risk, treatment response, or some combination of these criteria (Kendell 1989). However, the more limited information about whether a dimensional characterization accurately describes the patterning of data is nonetheless useful for purposes of guiding the specification of epidemiological investigation.

Without going into any details, the basis for taxometric methods is the derivation of predictions regarding empirical associations among symptom scores from models positing the existence of either discrete or continuous underlying constructs. All taxometric methods search for consistent inflection points in associations among the symptom measures in dimensional scales that would indicate the existence of a taxon rather than a latent continuum. A variety of strategies (Meehl and Yonce 1994) are used for this purpose and deal with cases where the symptom measures are discrete, continuous, or some combination of discrete and continuous. These same methods can also be used to provide rough guidance on where to place a caseness cut-point on a dimensional symptom scale to operationalize a taxon even in the absence of confirmatory biological tests (Meehl 1995). The simulation evidence is quite impressive in suggesting that these taxometric methods can sensitively discriminate multivariate symptom patterns that are generated by a taxon from those that are not and that, in the presence of a taxon, they can locate with good precision the cut-point on the dimensional symptom scale that optimally discriminates between cases and noncases. Based on these results, sociologists working with dimensional symptom scales could use taxometric methods to validate the assumption implicit in their work that there is no true discrete latent illness underlying the symptom reports.

I am unaware of any published studies that have attempted to discover a taxon underlying scores on screening scales of nonspecific psychological distress. However, my own preliminary analyses of several such scales suggest that no such taxon exists. This is not surprising in light of the fact that nonspecific distress screening scales are explicitly designed to screen for a wide range of distress syndromes. However, the evidence is quite strong for the existence of taxa in analyses of more specific symptom data. For example, taxometric studies provide strong evidence for a taxon underlying the symptoms of antisocial personality disorder (Harris, Rice, and Quinsey 1994) and for a separate taxon underlying the symptoms of schizotypy (Korfine and Lenzenweger 1995). Importantly, there is little evidence in these more focused taxometric studies for a taxon underlying screening scales of anxiety and depression. This means that current International Classification of Diseases and DSM conceptions of generalized anxiety disorder and major depression probably represent arbitrary (from the taxonic perspective) cut-points on syndromes that are actually continuous in the population. As described in the next section, I believe that it is still useful to think of generalized anxiety disorder and major depression and other nontaxonic syndromes in categorical as well as dimensional terms and that we can have nonarbitrary bases for deciding on cut-points to define these syndromes despite the failure to find a latent taxon. However, it is important to distinguish between situations where we can define a taxonic cut-point and cases where no taxon exists. It is also important to recognize that external criteria of a sort discussed in the next section are required to justify the selection of a cut-point in instances where the dimensional score is not taxonic.

Before turning to a consideration of categor-
ical assessment in the absence of a taxon, it is worth noting that taxometric research played only a minor part in the development of the diagnostic criteria in the current International Classification of Diseases and DSM systems. This was necessarily the case due to the fact that taxometric methods only came into widespread use in the 1980s and 1990s. The results of taxometric studies will likely be much more influential in the planned revisions of the DSM, which has been announced by the APA as having a planned release date of 2010, and of the International Classification of Diseases, shortly thereafter. Indeed, the groundwork is already being laid for using taxometric methods to help guide these revisions (Kessler forthcoming a). The World Health Organization has launched a major program of research known as the World Mental Health 2000 initiative to carry out nationally representative mental health surveys in 28 countries around the world, with a combined sample size in excess of 200,000 people (Kessler and Ustun 2000). One of the main goals of the World Mental Health 2000 surveys is to generate data that can be used to carry out taxometric analyses that will serve as an empirical foundation for the planned revisions of the International Classification of Diseases and DSM systems.

THE CASE FOR CATEGORICAL ASSESSMENT IN THE ABSENCE OF A TAXON

There are three important attractions of categorical assessment, even in the absence of evidence for a latent taxon. The first is that, from the perspective of the clinician, treatment decisions are often categorical. Blood pressure is measured dimensionally and has a monotonic relationship with risk of subsequent stroke. There appears to be no taxon for high blood pressure. However, clinicians have to decide when to intervene in this continuous distribution and this is usually a categorical decision. As a result, the National Heart, Lung, and Blood Institute has provided a categorical definition of hypertension (systolic blood pressure $\geq 140$ mmHg or diastolic blood pressure $\geq 90$ mmHg) in order to guide clinical categorical treatment decisions (NHLBI 1998). This cut-point was selected based on external considerations concerning epidemiological data on risks of heart attack and stroke associated with specific combinations of systolic and diastolic blood pressure rather than based on internal evidence for a taxon. In cases of this sort, where external criteria are used to define rational cut-points, it is reasonable to assume that the cut-points will change over time. Changes of this sort can occur based on considerations of cost-effectiveness (e.g., average treatment cost per year of life saved due to the treatment), or can be based on considerations of competing risks that compare the decrease in risk of morbidity and increase in estimated longevity created by treatment to the increase in risk of morbidity and decrease in estimated longevity created by other illnesses that might occur as an indirect result of the treatment (Gold et al. 1996). Political and financial considerations also come into play in making these changes.

The second attraction of categorical assessment is that it allows researchers to estimate prevalence of dimensional scores above the cut-points that are recognized as clinically significant. This is of considerable importance for needs assessment research, where some sort of threshold is required to determine how many people in the population are defined as in need of treatment. The definition of need can have dimensional characteristics, as when we discriminate among people classified as having severe disorders, serious but not severe disorders, and non-severe disorders. The definition can be influenced by moral as well as by scientific consideration and can consequently be made in different ways by different societies. However, this decision takes on a practical reality once it is made. As noted above, the impetus for the development of the first fully structured psychiatric diagnostic interview in the ECA Study came directly from this desire to estimate the number of untreated people in the community who might need treatment based on the definitions and criteria of the DSM-III.

It is important to realize that the existence of a taxon is not relevant here. Going back to the blood pressure example, it is important to know the number of people in the population who have untreated hypertension even though we recognize that the cut-point for defining hypertension is in some sense arbitrary. We also know that the policy significance of having this prevalence estimate is a good deal greater than the significance of knowing the
mean systolic and diastolic blood pressures of people in the population who are untreated. This is true even though interventions might focus on shifting the entire population blood pressure distribution downwards rather than reducing the proportion with hypertension (Rose 1992).

The same is true of mental disorders. Even though the syndromes of anxiety and depression are probably not taxonic, policy makers are much more interested in knowing the number of people with high scores on anxiety and depression symptom scales who are untreated than in having dimensional information. This is not to say that dimensional information is irrelevant. It would be interesting, for example, to examine the distributions of anxiety and depression symptom scales separately in community samples of people who are and who are not in mental health treatment, or to track entire distributions over time. Indeed, a case can be made that the best way to reduce the proportion of the population with scores so high that they are considered clinically significant is to develop intervention aims at shifting the entire population distribution downwards (Rose 1992). However, this way of thinking is still sufficiently uncommon that the critical questions asked by policy makers continue to involve the numbers and characteristics of the people at the upper tails of these distributions.

The third attraction of categorical assessment is that it makes it possible to obtain (usually retrospectively) information about lifetime occurrence. Dimensional measures, in comparison, limit assessment to recent time intervals such as the past week (e.g., Radloff 1977) or the past month (e.g., Derogotis 1983). The fact that categorical assessment makes it possible to speak about a discrete syndrome means that we can ask respondents about such things as how old they were when this syndrome first occurred and whether the syndrome has been episodic or chronic since that time. It is also possible to ask about such aspects as number of episodes, length of episodes, time between the end of one episode and the beginning of the next, and whether the length or time between episodes has become systematically shorter or longer over time. Information of this sort enriches our understanding of the natural history of these syndromes in a single survey much more than does the information obtained in dimensional measures on current symptom distributions.

THE CASE FOR DIMENSIONAL ASSESSMENT IN THE PRESENCE OF A TAXON

As briefly noted earlier in this paper, a case can be made for dimensional assessment even in the presence of a taxon. This is most clear in clinical research where, although treatment decisions are categorical, the evidence on which a treatment decision is based is almost always dimensional. How severe is the asthma? What are the scores on the dimensional liver function tests? In cases where the relevant dimensional measures are of borderline clinical significance, watchful waiting might be the appropriate initial clinical reaction even if the patient has the full set of symptoms that define the latent taxon. Treatment might be suggested if the dimensional scores are more elevated, while hospitalization might be the recommended course of action in even more extreme cases. Dimensional information is commonly used in this way in psychiatry. For example, cut-points for mild, moderate, and severe depression based on the dimensional Hamilton Rating Scale for Depression (Hamilton 1960) are routinely used to help guide clinical decisions regarding the use of medications and whether inpatient treatment might be needed.

A different attraction of dimensional assessment is associated with the recent development of brief, fully-structured diagnostic interviews aimed at providing a more rapid categorical assessment (e.g., 10–30 minutes) than the Diagnostic Interview Schedule or Composite International Diagnostic Interview (which typically take an hour or more to administer). These brief interviews generate diagnoses quickly by assessing only a subset of all the International Classification of Diseases and DSM criteria and making diagnoses based on best estimates rather than comprehensive assessments. As a result, the most accurate way to interpret the results of these assessments is in terms of probabilistic estimates of diagnoses. These probabilistic assessments represent a type of dimensional score, albeit a type that has a limited range requiring special analysis procedures, such as Tobit analysis, to be used when the scale scores are used as outcomes in regression analysis. The Composite International Diagnostic Interview Short Form scales, for example, were developed by abstracting subsets of items from the Composite International Diagnostic Interview
that were the strongest predictors of particular diagnoses in community surveys (Kessler et al. 1998a). Logistic regression analysis was then used to create rules for generating a series of disorder-specific dimensional scores in the 0 to 100 range for the predicted probabilities of having each disorder. While these dimensional scores are less precise than the dichotomous scores generated by the full Composite International Diagnostic Interview, the Composite International Diagnostic Interview Short Form can be administered in an average of eight minutes compared to approximately 70 minutes for the full Composite International Diagnostic Interview. The dimensional probability-of-caseness scores are adequate for many individual-level classification purposes. Furthermore, analyses aimed at studying the predictors of the categorical diagnoses underlying the dimensional scores can be carried out using conventional linear regression analysis in which the outcome variable is the log-odds transformation of the 0–100 scale. The linear regression coefficients obtained in this way can be interpreted as logistic regression coefficients (Kessler et al. 1999).

A similar sort of transformation can be used to reinterpret the dimensional scores obtained in standard screening scales of nonspecific psychological distress. However, in order to do this it is necessary to develop a meaningful categorical classification target and to show that the screening scale is strongly related to that target. A new 10-question screening scale of this sort was recently developed for use in the core of the redesigned U.S. National Health Interview Survey (Kessler et al. forthcoming). Like the early screening scales developed for use in the Midtown Manhattan and Stirling County studies, this new scale was designed to be a generate-purpose screen for broadly defined mental disorders, although the new scale aims to screen for disorders only in the anxiety-mood disorders spectrum. However, unlike the earlier scales, which were developed in the absence of clear diagnostic targets, the new scale was developed using modern psychometric methods with the explicit intent to screen for the prevalences of (1) any 12-month DSM-IV anxiety or mood disorder, (2) the subset of these disorders that meet criteria for a Serious Mental Illness stipulated in Public Law 102–321 (Kessler et al. 1996), and (3) the subset of Serious Mental Illness disorders that meet criteria for a Severe and Persistent Mental Illness proposed by the National Advisory Mental Health Council of the National Institute of Mental Health (NIMH; NAMHC 1993). Previous national surveys have estimated that the current general population prevalence of these three successively less common disorders are roughly 12 percent any disorders, 6 percent Serious Mental Illness, and 3 percent Severe and Persistent Mental Illness (Kessler et al. 1999).

Figure 1 shows Receiver Operating Characteristic curves for the relationships between scores on the 10-question dimensional screening scale and these three summary categorical diagnostic classifications. The sample on which these curves are based is the cross-validation subsample (n = 80) of a small clinical calibration sample that was enriched to include a high proportion of respondents with serious mental illnesses (Kessler et al. forthcoming). The clinical interviewers used to carry out the calibration interviews were all clinical psychologists who were carefully trained and monitored in the administration of the Structured Clinical Interview for Axis I DSM-IV Disorders (First et al. 1997), a semi-structured psychiatric diagnostic interview. Respondents self-administered the 10-question screening scale, followed by the Structured Clinical Interview for Axis I DSM-IV Disorders. Structured Clinical Interview for Axis I DSM-IV Disorders interviewers were blind to the respondent reports on the 10-question screening scale at the time the Structured Clinical Interview for Axis I DSM-IV Disorders was administered. The data were weighted to the population distribution of Serious Mental Illness prior to analysis because Receiver Operating Characteristic curves are sensitive to the oversampling of extreme scores. As shown in the figure, the areas under the Receiver Operating Characteristic curve, which can be interpreted as the probabilities that a random pair of respondents consisting of a case and noncase would be classified correctly based on differences in their dimensional scores (Hanley and McNeil 1982), show excellent classification accuracy for all three categorical outcomes. These Receiver Operating Characteristic curves can be used to transform screening scale scores into dimensional probability-of-caseness scores by using standard procedures (Peirce and Cornell 1993).
A different attraction of dimensional assessment exists for the risk factor researcher, where studies of the predictors of scores on dimensional symptom scales are much easier to carry out and offer much greater statistical power than analyses of the predictors of categorical transformations of these dimensional scales. As noted above, analysis of dimensional symptom scales makes sense even if there is a latent taxon underlying the dimensional scores. However, the usefulness of dimensional analysis hinges centrally on whether or not the risk factors are linearly related to scores on the dimensional scale. If the researcher wants his or her results to be relevant to the policy audience or to the clinical research audience, then the decision to focus on dimensional outcomes needs to be justified by demonstrating relevance at the clinical or policy threshold. In the absence of such a demonstration, potential users of the results have no way of excluding the possibility that significant predictors of the dimensional score are due to nonlinear effects of the predictors outside the clinical range.

I am aware of no sociological research on risk factors for dimensional measures of psychological distress that has systematically investigated the consistency of patterns of association, depending on whether the outcome is defined dimensionally or categorically. It is important to appreciate that the kind of analysis called for here is more complex than the conventional analysis of nonlinearity in the association between a continuous predictor and a continuous outcome variable in a linear regression equation. This conventional type of analysis evaluates the effects of polynomial transformations or discretized dummy variable transformations of the predictors in order to discover whether the predictors have nonlinear effects on the outcomes. This is inadequate because both the linear and nonlinear effects in such analyses are assumed to be constant across the range of the dimensional outcome. It is the latter assumption that needs to be tested if the results are to have clinical or policy relevance.

The analysis of the consistency of predictor effects across the range of dimensional outcomes is complicated by the fact that it requires an investigation of nested associations across

---

**FIGURE 1.** Receiver Operator Characteristic Curves for the U.S. National Health Interview Survey 10-question Screening Scale Predicting any 12-month Diagnostic and Statistical Manual of Mental Disorders—Fourth Edition (DSM-IV) Structured Clinical Interview for Axis I DSM-IV Disorder (Any), Serious Mental Illness (SMI), and Severe-Persistent Mental Illness (SPMI)*

* Based on analysis of n = 80 respondents in the cross-validation subsample of a community sample selected to validate the 10-question screening scale. For more details on the sample, see Kessler et al. (forthcoming).
Orthogonal transitions in the outcome dimension, typically using ordinal regression analysis (McCullagh 1980) or nested logistic regression analysis (Efron 1988). The logic of the latter, which can be implemented with conventional logistic regression software, is identical to the logic of discrete-time survival analysis in the study of temporal transitions (Willett and Singer 1993). The only difference is that we apply the analysis to symptom intensity rather than to time. A quick example can easily illustrate the logic. Consider the 10-question U.S. National Health Interview Survey screening scale of non-specific psychological distress described above. This scale, which is based on questions that have a five-category frequency response format (all of the time, most of the time, some of the time, a little of the time, and never), generates raw scores in a 0–40 severity range. A nested logistic analysis would begin by building a person-severity file that defined a separate dichotomous 0–1 outcome variable for each of the 40 pairwise severity comparisons embedded in this scale (0 versus greater than 0, 1 versus greater than 1, 2 versus greater than 2 . . . 39 versus 40). For example, the first dichotomous outcome would specify whether the respondent had a score of exactly 0 (coded 0) or greater than 0 (coded 1). The n observations, including a full array of predictors in addition to the score on this dichotomy and a dummy variable indicating that the observations defined the 0 versus 1–40 contrast, would then be stored in a special data file. The subsample of respondents with scores greater than 0 on the 0–40 scale would then be used to define a new dichotomous outcome variable that specified whether the respondent had a score of exactly 1 (coded 0) or greater than 1 (coded 1). This subsample of respondents would then be stacked into the same special data file with the outcome entered in the same field as the outcome in the original n records used to define the 0 vs. 1–40 contrast, along with a full array of predictors and a dummy variable indicating that the observations defined the 1 versus 2–40 contrast. The construction of nested contrast dichotomies would continue across the full range of the scale (e.g., a score of exactly 2 versus greater than 2 in the subsample of respondents with scores greater than one . . . a score of exactly 39 versus 40 in the subsample of respondents with scores greater than 38) until the stacked data file included all existing contrasts. Respondents with a score of 0 would contribute exactly one observation to the data file, while respondents with a score of 40 would contribute exactly 40 observations.

Logistic regression would then be used to analyze these data. The outcome would be the 0–1 dichotomy defining successive contrasts. Predictors would include the 39 separate dummy variables defining the 40 contrasts plus the substantive predictors. Instead of the 39 dummies, the 40 contrasts could be defined as a single continuous variable coded 1–40 or as a set of polynomials of this continuous variable or as a set of fewer than 39 dummies. Significance tests of the incremental improvement in fit of more refined specifications over the single 1–40 continuous variable could be used to test the assumption of linearity. The regression coefficients of the substantive predictors would be interpreted as the effects of the predictors on successive severity contrasts based on the assumption that these effects are consistent across the entire 0–40 scale range. Interactions of the contrast variables with the substantive predictors would then be used to evaluate the legitimacy of this consistency assumption. Conventional dimensional analysis would be justified only if it could be shown that the effects are, in fact, consistent across the full scale range.

An illustrative preliminary analysis using this nested logistic regression approach was carried out for the first 3,000 Part II respondents in the National Comorbidity Survey Replication (Kessler forthcoming b). Although National Comorbidity Survey Replication will eventually interview 10,000 respondents, only this subsample was available for analysis at the time of preparing this article. A total of 25,114 contrasts were observed on the 10-question screening scale for these 3,000 respondents. The results in the first column of Table 1 show metric linear regression coefficients based on unweighted data for the effects of sex (female coded 1, male 0), age (standardized to a mean of zero and variance of one), education (standardized to a mean of zero and variance of one), and race-ethnicity (dummy coded with “nonhispanic white” the deleted category compared to “Hispanic,” “nonhispanic black,” and “other”) in predicting scale scores in the n = 3,000 person-level data. Consistent with much previous research, the results show that scale scores are significantly higher among women than men, inversely related to age and education, and significantly higher among respondents defined as “other” race-ethnicity than “nonhispanic whites.” No attempt
<table>
<thead>
<tr>
<th>Persons Transitions</th>
<th>Subsample Transitions</th>
<th>Quintile 1</th>
<th>Quintile 2</th>
<th>Quintile 3</th>
<th>Quintile 4</th>
<th>Quintile 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>Female</td>
<td>1.8*(.3)</td>
<td>1.3* (1.2-1.4)</td>
<td>1.5* (1.3-1.8)</td>
<td>1.4* (1.2-1.8)</td>
<td>1.2* (1.1-1.5)</td>
<td>1.2* (1.0-1.5)</td>
</tr>
<tr>
<td>Age</td>
<td>-1.7*(.1)</td>
<td>.8* (.7-0.8)</td>
<td>.7* (.6-0.8)</td>
<td>.8* (0.7-0.8)</td>
<td>.8* (.7-.9)</td>
<td>.8* (.7-.9)</td>
</tr>
<tr>
<td>Education</td>
<td>-.5*(.1)</td>
<td>.9* (.9-1.0)</td>
<td>1.0 (.9-1.1)</td>
<td>1.0 (0.91.1)</td>
<td>9* (0.8-1.0)</td>
<td>.9* (.8-1.0)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>.2 (.5)</td>
<td>1.1 (.9-1.2)</td>
<td>.9 (.7-1.2)</td>
<td>1.3 (0.9-1.9)</td>
<td>1.0 (.8-1.3)</td>
<td>1.2 (0.9-1.6)</td>
</tr>
<tr>
<td>Black</td>
<td>.6 (.4)</td>
<td>1.1 (.9-1.2)</td>
<td>.9 (.7-1.2)</td>
<td>1.2 (0.9-1.7)</td>
<td>1.2 (.9-1.5)</td>
<td>1.0 (.7-1.3)</td>
</tr>
<tr>
<td>Other</td>
<td>2.4*(.3)</td>
<td>1.4* (1.3-1.6)</td>
<td>1.3* (1.0-1.6)</td>
<td>1.5* (1.2-1.9)</td>
<td>1.4* (1.2-1.8)</td>
<td>1.8* (1.4-2.2)</td>
</tr>
<tr>
<td>(n)</td>
<td>(3,000)</td>
<td>(25,114)</td>
<td>(5,528)</td>
<td>(4,306)</td>
<td>(6,048)</td>
<td>(4,463)</td>
</tr>
</tbody>
</table>

Note: *p < .05

a Based on analysis of the first n = 3,000 Part II respondents in the National Comorbidity Survey Replication (NCSR). For more details on the NCSR, see Kessler (forthcoming).

b SE represents standard error and OR (95% CI) represents odds ratio followed by the 95% confidence interval of the OR.
was made to elaborate the model to include substantively plausible specifications, such as a nonlinear effect of age (Newmann 1989) or an interaction between age and gender (Mirowsky 1996), because the data analysis was carried out merely for illustrative purposes.

The results in the second column of the table show logistic regression coefficients for the same predictors based on the n = 25,114 person-severity contrasts. These results are very similar to those of the person-level linear regression. This is not surprising in light of the fact that both specifications assume that the effects of the predictors are constant across the range of the outcome scale. The results in the third through seventh columns, however, show logistic regression coefficients for subsamples of the person-severity contrasts defined in terms of quintiles on the person-level 0–40 scale distribution. As shown in the table, there is remarkable consistency in the coefficients across the first four quintiles of the distribution, indicating that these predictors do, in fact, have consistent associations across the range of the distribution that defines minimal clinical significance of symptoms (i.e., the lower 80 percentile compared to the upper 20 percentile on the distribution). However, none of the predictors other than education is significantly related to symptom severity in the top quintile, which is the part of the distribution that taps the clinical range of symptom severity. Assuming that this conclusion continued to hold based on more fine-grained substantive analysis as well as sensitivity analysis, it would suggest that the principled way of analyzing these data dimensionally would be to truncate scale scores at the upper end of the fourth quintile before analysis and interpret results as providing information about the predictors of minimally clinically significant symptoms. It would also be possible to carry out a dimensional analysis of respondents in the highest quintile to study predictors of symptom severity in the clinical range.

THE INTEGRATION OF CATEGORICAL AND DIMENSIONAL ASSESSMENT IN EPIDEMIOLOGICAL RESEARCH

The diagnosticians who are involved in the revisions of the International Classification of Diseases and DSM systems are very interested in using taxometric assessment procedures to provide a principled empirical basis for making these revisions. However, biological psychiatric researchers have not shown nearly as much interest in the results of taxometric studies. The reason for this lack of interest should be clear from the foregoing discussion—that evidence for the existence or nonexistence of a taxon provides no guidance regarding the direction for future neurobiological research. Taxometric research provides only boundary information to support the view that an underlying biological entity is likely to exist. Biological psychiatrists do not need such evidence at this stage in the evolution of neuroscience. Taxometric research is very similar in this regard to population genetic twin research. The latter was the subject of great interest in the early days of biological psychiatry, when evidence documenting the heritability of psychiatric syndromes encouraged investment in neurobiological research by suggesting that there actually might be neurobiological bases of these disorders. However, as neurobiological research progressed and it became amply clear that there are powerful biological influences on mental disorders, the interest in population genetics faded because population genetic studies provide no help to molecular geneticists in their search for informative genes.

The same kind of low relevance of taxometrics for sociological and epidemiological research can be inferred from the fact that it makes sense to work with dimensional assessments of an illness even if it is possible to demonstrate that a taxon is likely to exist. This is true because people with the same illnesses can vary enormously in severity, and it is this difference in severity that occupies much of the attention of biomedical researchers. Similarly, it makes sense to work with categorical assessments even if it is possible to demonstrate that a taxon is unlikely to exist. This is true because many of the things that we conventionally refer to as illnesses represent extremes on dimensional measures of biological processes. Hypertension and Type 2 diabetes are good examples. The thresholds for where to define extremes as indicative of illness change over time as new knowledge about biological processes and treatments become available. It is important to understand why it is that some people have extreme values on these dimensions using analysis methods that do not assume linearity of effects across the full dimensional range.
These observations should make it clear that there are benefits of integrating categorical and dimensional assessments. One can see this kind of integration in clinical trials, where it is conventional to present results in terms of mean differences between treatment and control arms on a dimensional symptom severity scale and also in terms of proportions of treatment and control respondents with various categorically defined outcomes. The two most commonly reported categorical outcomes are "treatment response," defined as some specified magnitude of reduction (usually 50%) in the symptom severity score, and "recovery," defined as a reduction in the symptom severity score to some specified low point that is considered no longer clinically significant. It is not uncommon for these various outcomes to yield different results, such as showing that the treatment leads to a significant reduction in average symptom severity and to widespread improvement, but to little recovery.

One can easily imagine a similar mix of outcomes being useful in community surveys both to describe the distribution of mental health problems in the population and to study predictors and social consequences. Indeed, work along these lines is already beginning to appear in the literature. For example, there has been a great deal of interest over the past few years in the distribution and correlates of social phobia (Davidson 2000; Leclerc et al. 2000; Lépine and Pelissolo 2000; Montgomery 1999; Stein and Gorman 2001; Wittchen 2000) based on the finding in recent community surveys that a surprisingly high proportion of the population meets criteria for this disorder and the low proportion of these people who ever seek treatment (Chartier, Hazen, and Stein 1998; Davidson et al. 1993; Kessler, Stein, and Berglund 1998). This work has shown, though, that the mean of these community cases on standard dimensional scales of clinical severity is quite low and that only a minority of community cases can be classified categorically as having the subtype of social phobia that has been found in clinical studies to be associated with serious impairment (Stein, Walker, and Forde 1994). In addition, the predictors and social consequences of categorically defined social phobia and dimensionally defined social anxiety symptoms have been shown to differ in ways that shed light on the determinants of initial onset of the syndrome and determinants of increases in symptom severity (Kessler forthcoming c).

The World Health Organization World Mental Health 2000 initiative is the first large-scale psychiatric epidemiological study to attempt an integration of dimensional and categorical assessments across a wide range of syndromes. Like previous surveys in the ECA/National Comorbidity Survey tradition, the core of the World Mental Health 2000 surveys is a fully structured diagnostic interview, the Composite International Diagnostic Interview (World Health Organization 1997) that generates categorical diagnoses according to the definitions and criteria of both the International Classification of Diseases-10 and DSM-IV diagnostic systems. However, World Mental Health 2000 also includes two important kinds of dimensional assessments. The first is the 10-question National Health Interview Survey screening scale described above. This scale was included in World Mental Health 2000 so that nationally representative relationships of the dimensional scores with broad-gauged categorical diagnostic classifications based on the Composite International Diagnostic Interview could be generated along the lines of the results reported for the small calibration sample in Figure 1. The hope is that accurate benchmark calibrations of this sort will encourage the use of this screening scale in future surveys. Trends in the screening scale scores can be studied in such surveys both dimensionally (i.e., by charting trends in means) and categorically (i.e., by transforming scale scores to predicted probabilities of having Serious Mental Illness or Severe and Persistent Mental Illness and charting trends in the estimated prevalences of these categorical variables).

The second important kind of dimensional assessment used in the World Mental Health 2000 surveys involves a series of fully-structured versions of clinical disorder-specific symptom severity scales that are being administered to subsamples of respondents who meet criteria for Composite International Diagnostic Interview disorders. Included here, for example, are structured versions of the Multicenter Collaborative Panic Disorder Severity Scale to assess the severity of panic disorder (Shear et al. 1997), the Inventory for Depressive Symptomology to assess the severity of clinical depression (Rush, Carmody, and Reimetz 2000), and the Yale-Brown Obsessive
Analyses of the ECA and NCS data show that is especially true in light of the fact that post hoc surveys need to be supplemented with survey-based dimensional information on severity to be useful for policy purposes (Regier et al. 1998). This dimensional information is critical for understanding the severity found in clinical samples. We had no way of answering this question in the past because none of the dimensional symptom severity scales that are typically administered to patients are used in community surveys. The use of clinical disorder-specific symptom severity scales in the World Mental Health 2000 surveys will resolve this uncertainty.

Another important reason for including dimensional scales of this type in the World Mental Health 2000 surveys is that the high prevalences of International Classification of Diseases or DSM disorders in large-scale community epidemiological surveys based on fully structured diagnostic interviews like the Diagnostic Interview Schedule and the Composite International Diagnostic Interview (Kessler et al. 1994) and that close to one-fifth of respondents carry one or more DSM or International Classification of Diseases diagnoses at the time of these surveys (Kessler and Frank 1997). Clinical calibration studies have confirmed that these high prevalence estimates are consistent with those based on blind clinical interviews (Kessler et al. 1998b; Wittchen 1994). However, these high rates are so inconsistent with the expectations of clinicians that questions arise regarding whether the typical severity of these disorders is much lower than the severity found in clinical samples. We had no way of answering this question in the past because none of the dimensional symptom severity scales that are typically administered to patients are used in community surveys. The use of clinical disorder-specific symptom severity scales in the World Mental Health 2000 surveys will resolve this uncertainty.

SUMMARY

The decision by sociological researchers to use dimensional rather than categorical outcomes in studies of mental illness has traditionally rested on the twin assumptions that the associations of predictors with psychological distress syndromes are most accurately operationalized by using dimensional measures rather than categorical measures of these syndromes and that no true underlying discrete mental illnesses can reasonably be inferred to exist that would justify the creation of dichotomous transformations of dimensional measures. These assumptions have been asserted rather than tested. As described in this paper, methods now exist to test these assumptions. Such tests should be required to justify the use of dimensional assessments in future empirical studies. The critical test should be whether the predictors are consistently related to differences in symptom severity across the full relevant range of the dimensional distribution. If they are, then analysis of the dimensional version of the symptom scale makes most sense even if there is evidence of a taxon. If the predictors are not consistently related to differences in symptom severity, analysis of either a categorical, a truncated dimensional, or a nested set of categorical and dimensional transformations of the original dimensional scale make the most sense even in the absence of a taxon. Taxometric analysis, in comparison, while important in providing a principled basis for categorical distinctions, should not be decisive in determining whether to work with categorical outcomes. There are legitimate clinical and policy reasons for making categorical distinctions even in the absence of a taxon.
REFERENCES


Ronald C. Kessler is a Professor in the Department of Health Care Policy at the Harvard Medical School. He is the Principal Investigator of the National Comorbidity Survey (NCS) and of several follow-up replication surveys based on the NCS. He is also the Co-director of the World Health Organization World Mental Health 2000 surveys, a series of psychiatric epidemiologic surveys being carried out in over two dozen countries throughout the world. His research includes a wide variety of both naturalistic and experimental studies of psychosocial risk factors for, and consequences of, mental disorders.