Development and Psychopathology, **15** (2003), 529–551
Copyright © 2003 Cambridge University Press
Printed in the United States of America
DOI: 10.1017.S0954579403000282

Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology

ANDREW PICKLES[a] AND ADRIAN ANGOLD[b]
[a]*University of Manchester; and* [b]*Duke University Medical Center*

Abstract
The question of whether to view psychopathology as categorical or dimensional continues to provoke debate. We review the many facets of this argument. These include the pragmatics of measurement; the needs of clinical practice; our ability to distinguish categories from dimensions empirically; methods of analysis appropriate to each and how they relate; and the potential theoretical biases associated with each approach. We conclude that much of the debate is misconceived in that we do not observe pathology directly; rather, we observe its properties. The same pathology can have some properties that are most easily understood using a dimensional conceptualization while at the same time having other properties that are best understood categorically. We suggest replacing Meehl's analogy involving qualitatively distinct species with an alternative analogy with the "duality" of light, a phenomenon with both wave- and particle-like properties.

For many years a debate has raged over whether child and adolescent psychopathology should be regarded as consisting of a series of categorical phenomena (with individuals being either cases or noncases of various disorders) or as dimensions with psychopathology being just their negative extremes (Achenbach, 1966, 1985, 1991b; Sonuga–Barke, 1998). At the moment, the official nosologies, *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994); and the *International Classification of Diseases* (10th ed.; *ICD-10*); World Health Organization [WHO], 1994) and the research diagnostic interviews that implement them (Angold & Fisher, 1999; Shaffer, Fisher, & Lucas, 1999) fall firmly on the side of categorical diagnoses based on increasingly complex algorithms. However, for many purposes, questionnaires based on an explicitly dimensional conception of psychopathology, such as the Child Behavior Checklist (CBCL) and its congeners (e.g., Achenbach, 1991a, 1991c, 1992; Achenbach & Edelbrock, 1991; Achenbach & Rescorla, 2000) continue to be very widely used in both research and clinical practice. In this paper we revisit this debate from both theoretical and empirical perspectives and argue that both sides have been fighting under false colors because the questions at issue have been misframed. The central question is not "Is psychopathology scalar or categorical?" but "*Under what circumstances* does it make sense to regard psychopathology as being scalar and *under what circumstances* does it make sense to regard psychopathology as being categorical?" An essential part of our argument is that it is necessary to shift the debate away from trying to determine whether

This work was partially supported by Grants MH45268 and MH48604 to the first author and MH57761–04 and DA11301–05 to the second author from the US NIMH and NIDA.

Address correspondence and reprint requests to: Andrew Pickles, Biostatistics Group, School of Epidemiology and Health Science, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK. E-mail: andrew.pickles@man.ac.uk.

there are categorical states or dimensional levels of psychopathology toward considering the forms of *relationships with* other processes, either epidemiological or clinical. In both research and clinical practice it is these process-oriented issues that are usually of primary concern; and the form of these relatinoships, whether discrete or continuous, does not necessarily correspond to the supposed from of the psychopathology. For instance, even the most hardened categorialist will accept reductions in levels of symptomatology as evidence of treatment efficacy, regardless of whether study participants still meet criteria for suffering from the disorder at the end of the study. On the other side of the coin, it is hard not to imagine someone espousing a basically continuous view of aggressive behavior and refusing to acknowledge that it was important to decide categorically whether some individual was or was not a murderer. We will argue that most forms of psychopathology (indeed, most forms of pathology of any sort) manifest both continuous and discontinuous relationships with other phenomena. In coming to these conclusions, we intend to show that at every level, from the design of measures to the analysis of data, continuous and discontinuous functions are inextricably interwoven. To illustrate these points, we will use examples drawn mainly from two general population studies of children and adolescents: the Great Smoky Mountains Study (Costello et al., 1996) and the Virginia Twin Study of Adolescent Behavioral Development (VTSABD, Eaves et al., 1993).

A Brief Summary of the Debate Over Scales and Categories to Date

In the period following World War II through the 1960s, diagnostic categories for child and adolescent psychiatric disorders were defined only in the crudest terms (e.g., American Psychiatric Association, 1952, 1968) and most general population research was conducted with questionnaire measures of numerous specific behaviors or overall "disturbance" (Cullen & Boundy, 1966; Cummings, 1944; Gould, Wunsch–Hitzig, & Dohrenwend, 1981; Griffiths, 1952; Haggerty, 1925; Lapouse, 1966; Lapouse & Monk, 1958, 1964; Long, 1941; McFie, 1934; Olson, 1930; Srole, Langner, Mitchell, Opler, & Rennie, 1962; Wickman, 1928; Young–Masten, 1938; Yourman, 1932). Factor analytic studies began to appear in the 1940s. These formed the basis for what later emerged as a fairly consistent set of factors resulting from parent-report questionnaires. (See Achenbach & Edelbrock, 1978, for a scholarly summary of the earlier work.) For most psychopathology research, "diagnosis," insofar as it was considered at all, was defined in terms of scoring above some percentile on the particular scale employed. Wilson (1993) argued that this dimensional view blurred the distinction between normal and abnormal and, being associated with a plethora of unsubstantiated theorizing about psychosocial causes of mental ill-health, contributed to a breakdown of clinical consensus and to the low professional status of psychiatry within US medicine. While this is probably laying too much blame for the ills of psychiatry at the door of dimensional models, these approaches certainly led to wildly varying estimates of the prevalence of psychopatholgy in children (Gould et al., 1981). However, the factor analytic tradition proved capable of generating some highly reliable, replicable, and internationally reproducible dimensions of psychopathology (Achenbach, Conners, Quay, Verhulst, & Howell, 1989; Achenbach & Edelbrock, 1978; Crijnen, Achenbach, & Verhulst, 1997).

By 1980 the basis for the categorical approach had been substantially strengthened. The first major diagnostic general population study of child and adolescent psychiatric disorders was conducted on the Isle of Wight during the 1960s. Similar methods were then used in a second study in an inner city London borough and in a follow-up on the Isle of Wight (Berger, Yule, & Rutter, 1975; Graham & Rutter, 1973; Rutter, 1965, 1976; Rutter, Graham, Chadwick, & Yule, 1976; Rutter, Tizard, & Whitmore, 1970; Rutter, Yule, Morton, & Bagley, 1975; Rutter, Yule, & Berger, 1974; Rutter, Yule, & Quinton, 1974; Yule, Berger, Rutter, & Yule, 1975). Second, Rutter and colleagues developed a multiaxial classification scheme that resulted in the WHO publishing an addendum to the *ICD-9* on the classification of child and adolescent psychiatric disorders (Rutter et al., 1969; Rutter, Shaffer, & Sturge, 1979). These studies and

the diagnostic scheme upon which they depended showed that categorical child psychiatric diagnoses were both feasible and capable of yielding results that were scientifically interesting and valuable for planning purposes. They contributed to the neo-Kraepelinian synthesis that was moving to dominate US psychiatry, in which mental illness "consist(s) of a finite number of disease entities, each with a distinct pattern of symptoms and course, and with distinct causes, treatments and neuropathologies" (Kendall, 1991, p. 1). That domination reached fruition, of course, in the third edition of the *DSM* (*DSM-III;* American Psychiatric Association, 1980). However, in order to cover a full range of child symptomatology, the *DSM-III* defined a variety of diagnostic categories for which there was little available validation (Rutter & Shaffer, 1980). Despite its many limitations, and much trenchant criticism of weaknesses in its scientific underpinnings (see, e.g., Blashfield, 1982), the basic approach adopted by the *DSM* has been incorporated into the *ICD-10* (Taylor, 1994; WHO, 1993, 1994), and now provides the leading paradigm for child and adolescent psychiatric research.

Clarifying Some Rhetorical Terminology

What is "empirical?"

In discussions of the relative merits of *DSM*-type diagnostic categories and symptom scale-based approaches to psychopathology, the latter are sometimes referred to as being "empirical" or "emprically derived" (see, e.g., Achenbach, 1985). The implication seems to be that diagnostic categories are *not* empirically derived. Because science is substantially an empirical enterprise, this is tantamount to suggesting that such categories are unscientific. The *Oxford English Dictionary* defines the relevant uses of the term empirical as follows: "Of a physician: that bases his methods of practice on the results of observation and experiment, not on scientific theory; Pertaining to or derived from experience." Note that empirical does not exclude but is not restricted to meaning "derived by principal components analysis with varimax rotation" (to summarize the basic scale-based approach). The group consensus methods used to develop current child psychiatric nosologies like the *DSM-IV* are firmly based on clinical observation and current research, and are, therefore, empirical. For instance, the substantial changes made in the diagnostic criteria for oppositional disorder and the anxiety disorders in childhood and adolescence in *DSM-IV* were based upon a range of research observations, including field trials set up specifically to examine the effects of a variety of possible changes to the criteria. The point here is not that the resulting changes were necessarily correct, or led to the identification of definitively distinct disorders, but that empirical methods are now equally characteristic of the development of categorical nosologies and dimensional approaches for developmental psychopathology. Of course, it remains a problem to decide *which* empirical approach will best advance understanding of a given question.

Quantitative versus what?

In comparing assessment approaches, Achenbach (1985) also contrasts the use of quantitative measures with categorical diagnoses, going on to explain that by quantitative measures he means checklist scores as opposed to *DSM* diagnoses. While his critique of the lack of definition of items in the *DSM-III-R* categories has a great deal of merit, the more general undertone identifying science with quantification is being used implicitly to devalue categorical approaches. Many real quantities are genuinely binary or polychotomous and not continuous: for instance, one either dies or one does not; one is either homozygous or heterozygous (e.g., for phenylketonuria or Huntingdon disease). The argument should be about which metric is appropriate and whether the measure used is adequate to the task of placing individuals in proper relation to one another on that metric.

The Construction of Scales and Categories

The role of expert opinion

An often repeated argument in favor of dimensional measurement is that real (numerically quantified) associations between phenomena are the basis of scale score based syndrome

descriptions, whereas categorical nosologies are dependent upon expert opinion. There are indeed differences between the dimensional and categorical approaches along these lines, but they are not as extreme as some would have us believe. In the development of a symptom scale, only a limited number of items are ever included in the item pool for analysis. Who chooses these items? The scale developer does, of course, who supposedly deserves the title of "expert." Sometimes the scale developer will poll others to assist in the definition of scale content, but those others will usually be clinicians (just as in the case of categorical diagnosis; see, e.g., Achenbach, 1966).

Expert decisions also enter the process in the analysis stage. A recurring theme in the measurement of psychopathology is whether to include or exclude those rare but characteristic symptoms of a disorder in the item pool, a decision that is often coupled to the decision as to whether the focus is on measurement of clinical or general population samples. Without special adaptation of the instrument, and particularly in general population samples, such items can often appear to degrade the measurement performance. As a consequence of such preliminary analysis, the initial item pool of the CBCL was winnowed down to exclude items that were rarely reported by parents as being positive. The decision to exclude such items obviously has a bearing on what the final item content of each factor observed will be.

The choice of analytical technique can also have an effect on the content of the dimensions resulting from a dimensional measure. The apparent need to reject rare items is commonly exaggerated by the use of ordinary factor analysis, a circumstance in which the use of formal item response models is clearly more appropriate. Many general checklist developers (Conners, 1997; Quay, 1977; Verhulst & Achenbach, 1995) used principal components analysis with varimax rotations to derive factor structures to provide internal validation. What would the results have been had they decided to use maximum likelihood factorization, oblique instead of orthogonal rotations, different rules for factor retention, or different factor loading criteria to decide which items would count toward final factor scores?

The point here is not that the dimensional measures we have are defective, but that their contents are not the result of simply "finding out" the structure of things in the real world. At every level, the things that are "discovered" are defined and constrained by their developers. Moreover, once discovered, they became reified, constraining subsequent thought and observation. Of course, exactly the same is true of categorical nosologies. Diagnostic interviews tend to focus on measuring only the phenomena mentioned in the current diagnostic criteria (and tend to be modified if the criteria change). Thus, the diagnosis, once defined, becomes reified because noncriterial symptomatology is no longer measured. For instance, the ways in which the category of depression has been implemented in the *DSM* system is inimical to demonstrating possible differences between child and adult manifestations of depression. On the other hand, the use of clinician and research experts to revisit and refine official diagnostic criteria means that a process is in place to implement warranted changes. The fact that clinicians are also involved in these reviews offers a particular opportunity for refining criteria in the light of extensive clinical experience, without the need to wait while new instrumentation is developed. This is far from being a perfect system, but it has proved quite capable of avoiding ossification. Indeed, a common complaint is that the American Psychiatric Association changes the DSM criteria too often.

Common measurement practices and assumptions

Cairns and Green (1979) outlined a number of assumptions underlying the use of rating scales, which, it turns out, also underlie the use of diagnostic criteria. First, consider the *DSM-IV* diagnosis of oppositional defiant disorder. Eight symptoms are to be considered, and four must be present in order for the diagnosis to be given. The second criterion is "often argues with adults." It would seem that the clinician (or computer diagnostic algorithm) must make several judgments in order to determine whether this criterion is met: (a) Does the child manifest the behavior "arguing with adults?" (b) How often does the child mani-

fest that behavior? (c) Is that frequency great enough to be called "often?" The second of these questions involves a dimension, and the third, the imposition of a "cut-point" on that dimension. All the criteria for oppositional disorder involve the same basic format. That is, all of them require the diagnostician to jump back and forth between categorical judgments (such as "does the child argue with adults—yes or no") and dimensional judgments (such as "how often"). Once all the criterion symptoms have been assessed, the number of positives must be counted; if their sum is four or more, then the diagnosis is given. Once again, a dimension (number of symptoms) is being constructed and then reduced to a category by means of a cut-point.

Now, consider item 3 of the symptom section of the CBCL, "argues a lot." This time it is the parent who must make the categorical decision "does my child argue?" Next she must consider how often the child argues (dimensional) and then decide whether that is "a lot" (categorical). The final stage involves a 3-point choice deciding whether the result of the earlier deliberations should result in a final answer of *not true*, *somewhat or sometimes true*, or *very true*. This last involves a shift back into dimensional mode, with the minimum number of levels to avoid being a categorical decision. We have been hard put to come up with any examples of symptoms that do not involve this sort of back and forth.

Second, it must be assumed that the informant shares with the diagnostician or scale developer an understanding of exactly which behaviors of the child represent the attribute of interest. However, it is obvious to any clinician that it often requires hard work to find out what you want to know because nonclinicians do not all use the same psychopathological terms in the same way. It is also obvious to anyone who teaches clinicians that they do not all share the same definition of every symptom. Neither the *DSM-IV* nor any checklist that we know of provides definitions of symptom items. Consider CBCL item 5, "behaves like opposite sex." One can hardly expect that everyone has the same notion of what "behaving like the opposite sex" entails. Exactly the same problem arises with the criteria for *DSM-IV* gender identity disorder. We doubt that any two clinicians will agree on exactly what constitutes "intense desire to participate in the stereotypical games and pastimes of the other sex." However, interviewer-based interviews have gone some way toward providing (operational and/or conceptual) definitions for interviewers and clinicians in an attempt to improve standardization at the symptom level (Angold & Fisher, 1999).

Third, the informant must be able to extract the relevant behaviors or states from the stream of everyday life and determine how often they occur. We would also add that this must be done in relation to the relevant time frame (e.g., the past 6 months for the CBCL and a variety of frames for *DSM-IV* diagnoses).

Fourth, the informant or diagnostician must then reduce the information already extracted to the appropriate metric for the final coding (e.g., *not true*, *somewhat or sometimes true*, *very true*, or *often true* on the CBCL or *symptom present/absent* for *DSM-IV*). Different parents judge the frequencies necessary to fall into such categories very differently. It is also worth noting that there is very little information about what constitutes normative behavior as far as most symptoms are concerned. Until recently (Angold & Costello, 1996), for instance, there have been, as far as we know, no data on how often oppositional disorder symptoms occur in the general population. In other words, the decision as to where in the frequency distribution to set the cut-point for "often argues with adults" has necessarily been left to the vagaries of individual guesswork. It would seem, therefore, that the measurement processes that are used to obtain categorical and dimensional characterisations of psychopathology share rather more in common than the proponents of each would have us believe.

Does it Matter?

Intervention

In certain circumstances, decisions must be made about intervention, and for clinical treatment these decisions are invariably categorical. For instance, before treating a child with stimulants, it is necessary to determine whether that child has symptoms of sufficient intensity to warrant such treatment. The *DSM-IV* cate-

gory of attention-deficit/hyperactivity disorder (ADHD) defines a group of children who are likely to benefit from such treatment. However, it is not the case that someone with a minimal amount of ADHD symptomatology will benefit from a minimal amount of stimulant medication. Rather, the decision to prescribe stimulants should institute a full trial of stimulants in reasonable doses. It does not matter for the purposes of our argument *how* the decision to provide treatment is made. It could be argued that such a decision should be based on the results of a well-known questionnaire (such as the Conners scale in the case of ADHD) or even on neuropsychological testing, without recourse to the *DSM-IV* criteria. But whatever assessment method is used, some cutoff point will have to be used to determine whether to institute treatment. Thus, no matter how dimensional the approach used for assessment, at the point at which a decision to treat or not to treat is made, all the assessment information must be reduced to a cateogical statement. It is usual to call the categorical statement a "diagnosis."

This view reflects the clinical perspective. Were we to adopt a community health perspective, then interventions need not necessarily be targeted at only those currently with the disorder. A dimensional view would suggest that health benefits can be obtained by reducing a subject's score, wherever they are on the distribution of scores, or that at least a downward shift in the mean would also deliver a reduction in the prevalence of the pathological upper tail. From such a perspective quite a different set of interventions become eligible for consideration (Offord, Kraemer, Kazdin, Jensen, & Harrington, 1998), although few of them would involve mass pharmacotherapy. Of course, community health interventions are also possible for disorders viewed as categorical, but these commonly involve targeting effort through the use of screening. Thus, the conceptualization and measurement of psychopathology as either dimensional or categorical may have both intended theoretical implications for and unintended force of habit associations with clinical practice and public health policy.

Communication

Having an agreed system of measurement is essential for both scientific advance and for providing a communication bridge between laboratory and clinical settings (Sonuga–Barke, 1998). However, the choice between categories or dimensions for psychiatric outcomes tends also to influence how risk and protective factors are measured and which specific tools for analysis and interpretation are used. Categorical outcomes tend to be associated with categorical predictors and with methods of analysis like logistic regression and log-linear models that report effects in terms of odds ratios. Dimensional outcomes are commonly associated with dimensional predictors and analyzed using conventional analysis of variance and regression that report effects in terms of mean differences, sums of squares, partial correlations, and proportion of explained variance (r^2). Although there are notable exceptions, to a considerable extent, psychiatry does not have its own training schemes in research methodology. Instead, the traditional contributing disciplines of psychology and epidemiology present independent methodological frameworks for dimensional and for categorical outcomes. Moreover, statistical texts traditionally place methods for dimensional and categorical data in different chapters with little or no attempt at integration. As a consequence, although we may be able to undertake and interpret analyses within each measurement tradition, how many of us have any feel for how the effects described in one tradition map onto effects described in the other? We suspect our skills and knowledge in this respect are very poor.

Consider the case in which a psychologically trained researcher regresses a normally distributed risk on a normally distributed outcome while our medically trained researcher imposes cut-points on both of these continuous measures, cross-tabulating the resulting binary risk and binary outcome in order to estimate an odds ratio. Figure 1 shows the results of a set of simulations of the relationship of the psychologist's r^2 value to the medic's odds ratio for different choices of cut-point. We see that for any one set of cut-points, a sim-

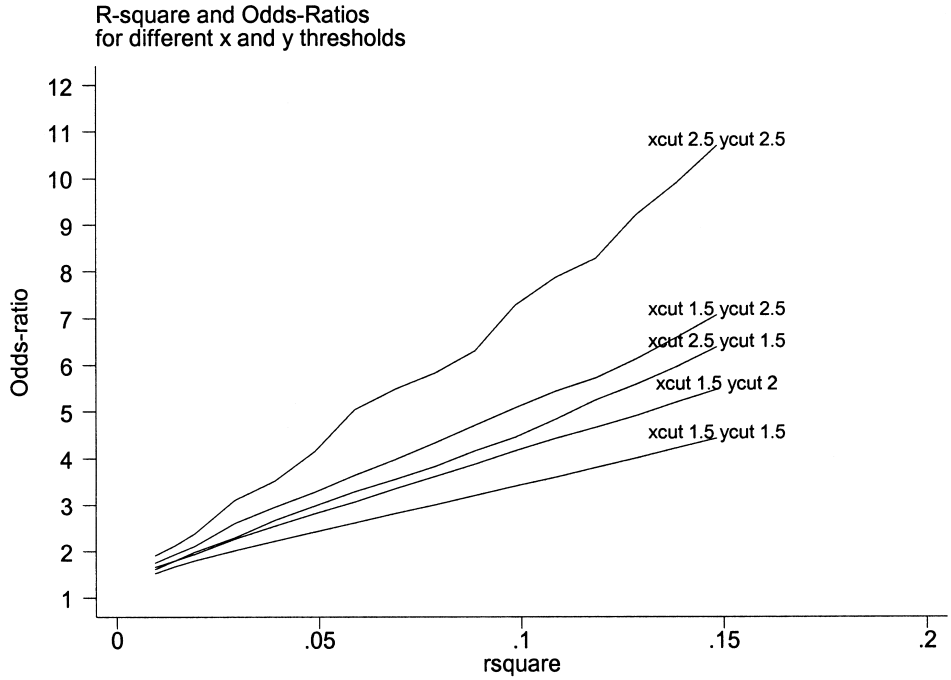

Figure 1. The relationship between dimensional and categorical effect estimates.

ple linear relationship prevails between the medic's odds ratio and the psychologist's r^2. For example, setting both cut-points at 1.5 *SD* identifies 6.7% as exposed to the risk and 6.7% as exhibiting the outcome, and the estimated odds ratio turns out to be about 35 times the r^2 value. Unfortunately, different choices of cut-point for either the risk measure or the outcome measure lead to different linear relationships. Cut-points of 2 and 2.5 *SD* identify 2.3 and 0.6% of the continuous distribution to be in the positive category, respectively. Raise the cut-point of either risk exposure or outcome to 2.5 *SD* (or raise both of them to 2 *SD*) and the odds ratio becomes about 50 times the r^2 value. Raise both cut-points to 2.5 *SD* to examine how extreme risk exposure is related to extreme outcome, and the estimated odds ratio will be more than 80 times the r^2 value of the underlying continuous measures. This figure tells us that to translate the reports of effects from the language of categories to the language of dimensions; although it maybe feasible in simple cases, is not entirely straightforward and is likely to be difficult for more complex problems.

There are some particular models and settings in which natural relationships arise. Where we have exposure to risk measured on a continuous scale, the case-control methodology leads us to expect that the cases will have a higher mean exposure than controls. Where the individual exposure measures are normally distributed around their respective group means and they have a common variance, then there is a little known but the simple relationship in which the difference in means divided by the within group variance turns out to be equal to the log-odds ratio for the effect of the risk measure on outcome (Pickles & Clayton, 2002). This equivalence is valuable not only conceptually but also practically for power calculations and more advanced modeling.

The so-called normal liability model also provides a framework that is helpful for communicating in the language of both perspectives. As in our simulations above, a categorical outcome can be viewed as having been obtained by placing a threshold on a potentially continuously scored outcome. However, in the normal liability model the analysis is

explicitly based on the assumption of the normality (conditional on predictors). This leads to the estimation of effects on the probit scale, rather than on the logistic (log-odds ratio) scale of traditional epidemiology. The advantage is that, subject to the usual vagaries of sampling error and the correctness of the normality and linearity assumptions of the model, the estimated effect of some predictor that we might obtain from a probit model of the categorical outcome is the same as the regression coefficient from a linear regression analysis of the dimensional outcome. It also provides a single framework within which simultaneous analysis of categorical and dimensional outcomes can be undertaken and for the application of latent variable models to such data (e.g., Muthén & Muthén, 2001). It would seem, therefore, that studies could be more easily compared and understood were we to use and report results from such probit and regression based models rather than using logistic regression.

Unfortunately, although the probit design offers advantages with respect to comparability for the categorical–dimensional issue, it has disadvantages when it comes to comparability over sample design. Sample selection has been crucial to scientific advance in epidemiology, with heavy oversampling of cases in case-control designs and heavy oversampling of the risk exposed in high-risk designs being key elements of many powerful studies. As a measure of effect, the odds ratio is the only measure that is unaffected by the use of one or other of these designs and gives (within sampling error) the same estimate as if a simple random sample of a general population cohort had been used. Systematic review across an immense range of studies is therefore possible. By contrast, effect estimates on the probit scale do not share this invariance property and will vary with the design of the study. The relative sizes of coefficients (comparing one risk to another) and their significance will typically be very similar, whether using either probit or logistic, but only for a quite limited range of outcome rate is there a simple relationship between the absolute size of the coefficients (the log-odds coefficient being about 3 times the probit estimate within that range, Maddala, 1983). Outside of this range, translation becomes more difficult.

This problem of translating from the language of categories to the language of dimensions is an important one, and it makes productive multidisciplinary research more difficult to achieve. Within research groups it suggests that we should be willing to run parallel analyses where we can. In research dissemination it suggests that, at the very least, we should provide additional pieces of information to enable results to be translated from one perspective to the other, and we may need to communicate key results in both languages. This would require authors and editors to adopt a more flexible use of journal space (article text, footnotes, appendices, and supporting web-based material). For methodologists it suggests that we should be doing more not only in providing integrated training but also in conducting more pragmatic research to chart in more detail areas of equivalence and difference and elaborate measures of effect and impact more generally (Kraemer et al., 1999).

Metatheoretical considerations

Sonuga–Barke (1998) has made much of the metatheoretical implications of adopting a categorical or a dimensional approach. He argues that the dimensional view is more open to environmental explanations of pathology, notably psychosocial explanations. By contrast a categorical view lends itself to within individual explanations, notably neurobiological explanations. We suspect that this association between measurement approach and type of explanation is just a historical accident resulting from the coincidental timing of interest in diagnostic categories and biological psychiatry, on the one hand, and guild issues of psychologist versus psychiatrist, on the other. As far as we are concerned there is nothing intrinsic to categories or dimensions that predisposes to explanations involving either nature or nurture. Nonetheless, we would agree that there are features about categorical or dimensional measurement that make them more or less suitable for operationalizing particular types of theory.

The categorical view lends itself to explor-

ations of interactions and more fully multivariate analyses, such as log-linear models and developments thereof. Those wishing to operationalize interactionist theory (Magnusson, 1988a, 1988b), have found this much easier to do within a categorical framework, using techniques such as configural frequency analysis (von Eye, 1990). Indeed, in some circles, the dimensional approach has become almost synonymous with "variable-based analysis," a pejorative term now used by those who prefer a categorical "person-based analysis." However, the source measures for such analyses need not be categorical but can be made categorical through the use of cluster analysis (e.g., Bergman, 2001). Moreover, some variable-based models such as random effects growth-curve models can capture and display some of the key features of the interactionist view, notably individual differences (see Pickles, 1989); and structural equation models are slowly becoming more elaborate, with mixture models (Muthen & Muthen, 2000), nonparametric discrete class factors (Rabe–Hesketh, Pickles, & Skrondal, 2002), multiplicative random effects (Pickles et al., 1996, 1998), and interactions and nonlinear effects (Schumacker & Marcoulides, 1998) now all being possible. Analyses that reflect the interests of interactionists are, therefore, becoming possible in a dimensional framework. However, much remains to be done both conceptually and in improving software implementations.

Categorical and dimensional views also lend themselves to the consideration of rather different types of developmental mechanisms. For instance, a substantial body of psychological theory has regarded development as progression through a series of stages (e.g., Kohlberg, 1976; Piaget, 1932, on moral development). The supposition of a stage, in fact, provides a model with a remarkable range of capacities. First elaborated in the context of cancer development these have also been described in the context of developmental psychopathology (Pickles, 1993). A key supposition of such theories is that each stage is qualitatively distinct from the one before and from the one that follows. Stage progression may then provide an opportunity for developmental changes in etiology, with one set of risk and protective factors influencing progress from stage A to stage B, whereas another set is important in the progression from stage B to stage C. The manner in which the effects of risk and protective factors combine to increase or decrease the rate of the final outcome will depend upon the transitions they impact upon. If the factors operate on the same stage transition, the null expectation is that they represent different "causes" or "pathways" and thus are likely to have effects that combine in an additive fashion. By contrast, factors that act on stage transitions at different points in the sequence can be expected to have effects that combine multiplicatively: the effect of the first factor increases the potential pool of subjects available on which the second factor can operate. Thus, how the effects of factors combine is potentially informative as to the structure of stages. Positing fewer or more stages also influences the expected age distribution at which the final developmental stage would be reached, and even the forms of pathology that we might entertain, for example, suggesting forms of pathology consistent with halted, delayed, or premature progression. It is also possible to operationalize stages as latent classes, which are either not directly or only partially observable (Macready & Drayton, 1994).

In a similar fashion, developmental pathways provide a means of linking potentially theoretically distinct steps into a chain of simple transitions. Such pathways have considerable intuitive appeal; and, when combined with bifurcating graphical displays, they are capable of conveying valuable information with respect to both the absolute and relative impacts of risk factors. In Figure 2 (adapted from Hill et al., 2001), the path thicknesses allow a comparison of both the relative and the absolute frequencies and the co-occurrence of recalled child sexual abuse and neglectful parenting and their association with gaining supportive adult love relationships. The position and angles of the paths shows the individual and joint impacts of the factors on depression during adulthood, including the effect modification exhibited by the protective impact of a supportive adult love relationship being restricted to those experiencing neglectful parenting, not childhood sexual abuse. This simple diagram is thus ca-

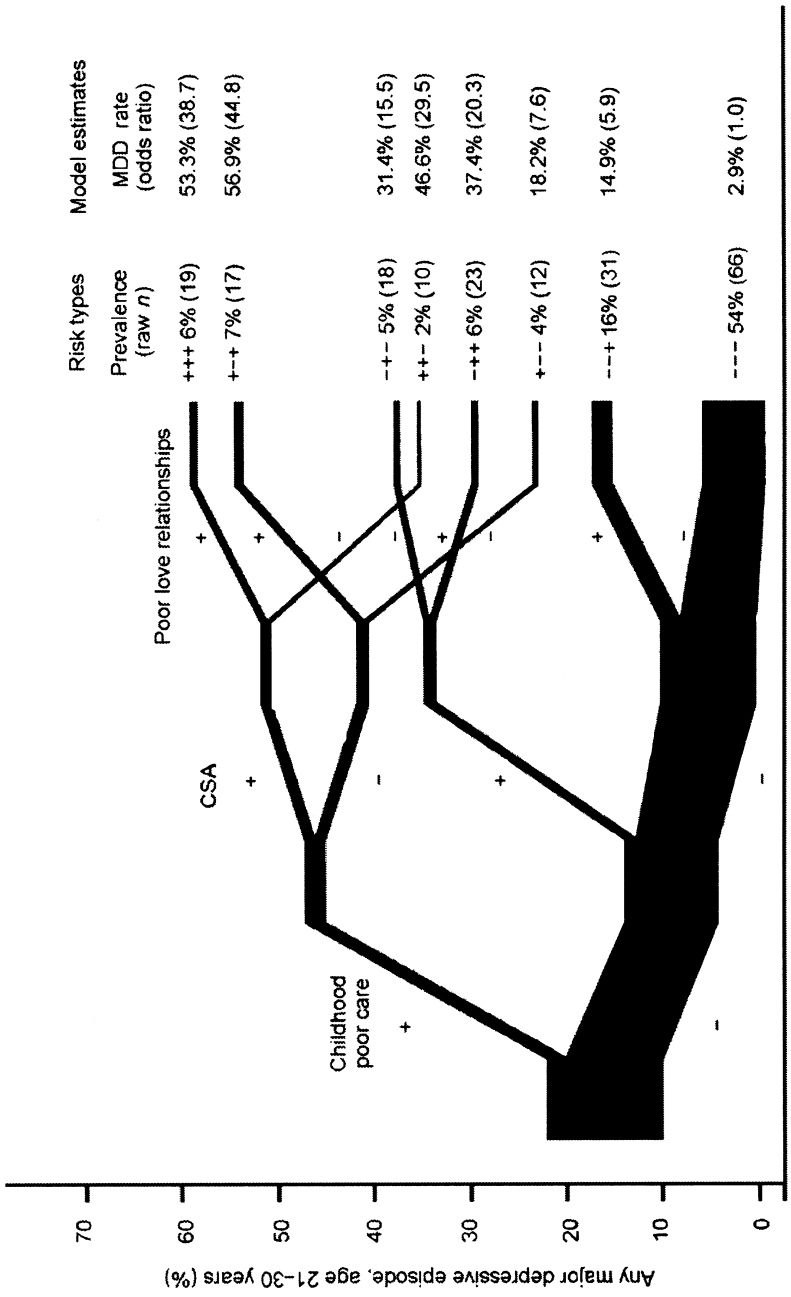

Figure 2. A graphical display of categorically defined developmental pathways (Hill et al., 2001).

pable of displaying complex multivariate relationships linked to clearly articulated theory. We know of no way of constructing a comparable diagram where the risks are treated as continuous variables.

This is not to say that mechanisms cannot be analyzed and displayed using models and methods based on dimensions. However, perhaps unlike physicists, we do not seem to have developed our intuitive grasp of such models very far. Formal dynamic models for continuous variables are typically very much more demanding of mathematics and require graphical display that can seem substantially more abstract than those described above. If we are willing to make wide-ranging linearity assumptions, then path diagrams from the structural equation modeling tradition can provide highly effective and parsimonious representations of complex multivariate problems. However, the difficulties in examining interactions and nonlinearities and the lack of highlighting of individual differences are substantial limitations of such methods. The language of categories seems to have a simplicity of vocabulary and grammar that is more sympathetic to our developmental theories. However, we should not allow this to prevent us from continuing to consider alternative models based on dimensions and improving our fluency with the language of continuous variables suitable for their description.

Relevant Forms of Evidence for Categories or Dimensions

Bimodality

Can we tell from the data we collect whether the scientific reality is one of dimensions or categories? Some phenomena relevant to psychopathology are self-evidently categorical or so nearly so as to be reasonably regarded as being categorical under most circumstances: sex comes to mind as an example. A phenomenon might also be so dramatically bimodally distributed that it makes little sense to treat it in any way but categorically, except in studies that concentrate specifically on the rare individuals who fall between the two modes.

Are we able to reveal categories more generally from the scores on a "dimensional" scale? The often cited bump at the lower end of the otherwise normal IQ distribution is an indication of the presence of a group of individuals with a range of disorders affecting IQ that are rarely found in the rest of the IQ range. A number of attempts to detect points of rarity and humps in symptom distributions have been made. That these have mostly been unsuccessful is exactly what the statistical theory on the monotonic transformation of continuous distributions would have led us to expect, because this makes clear that any continuous distribution, however bimodal, can be mapped into a unimodal one without disturbing the rank order of subjects. Only if we have other theory and evidence that justifies the particular level of measurement scale used and thus rules out the use of such a transformation can bimodality be used as evidence of a category. In fact, distributions of symptom counts in both children and adults appear, besides the discretization arising from consideration of a finite pool of symptoms, to be resolutely continuous.

Skewed distributions

Can we at least say that pathology is the tail of a skewed rather than a normal distribution? The form of any distribution of psychopathology will depend on the way in which psychopathology is measured. If one includes items in one's symptom scale that have varying prevalence (one from another) and reflect varying degrees of normality and abnormality, then it is easy enough to generate a roughly normally distributed curve. However, most of what clinicians would regard as being symptoms are simply absent in most people, with the result that general population symptom scores from interviewer-based interviews and many questionnaire item totals are heavily skewed to the right (i.e., most people have zero or very low scores). Such item total distributions should not be analyzed using standard normal theory regression's but they can be analyzed directly by the use of transformations, robust parameter covariance estimation, or an appropriate choice of generalized linear model based on Poisson, gamma, or inverse

power distributions. An alternative approach is to adopt an explicit measurement model in which the items are considered as measuring some underlying liability or propensity but are also subject to the impact of other factors, loosely referred to as measurement error. In their seminal work on psychometric scale construction, Lord and Novick (1968) showed that highly skewed item totals are a natural consequence of the use of low prevalence items even when the true liability distribution is normal. Using antisocial behavior and depression items from the National Longitudinal Study of Adolescent Health and the National Longitudinal Survey of Youth, van den Oord, Pickles, and Waldman (2003) compared the fit of models that assumed normal and skewed liability distributions. They found little evidence to reject the assumption of a normal liability distribution.

Latent classes and nonparametric maximum likelihood

Does the foregoing evidence demonstrate that psychopathology is, in fact, continuous and possibly normally distributed? This is not necessarily so. Just because a distribution is continuous does not mean that the phenomenon underlying it is not categorical. Suppose for a moment that there really were a brain disease called depression and you either have it or do not. Suppose also that at some point in the future some aspects of the mechanism of this disease will be discovered so that an accurate diagnostic test will be available; but, for now, we have to rely on asking a lot of questions about phenomena that are related to the real disease but also have a range of other causes. Let us also assume that the disease is not very common (say it affects 4% of the population) but some of the other causes of individual "depressive" symptoms (e.g., primary sleep disorders, anxiety disorders, bereavement, physical illnesses) are as common or more common. Let us also take into account that our available questioning techniques are imperfect measures, even at the symptom level. In other words, we face measurement error at both the symptom and diagnostic levels. What would we expect the distribution of "depressive" symptoms to look like? The result would be a distribution in which many people had a few symptoms and a few had many symptoms, with no sharp cutoff between the two. Our "real" depressives would be concentrated in the upper tail of the distribution, but because of the imperfections of our question-based assessment approach, some would be in the lower body of the distribution. Thus, even if we were measuring the correct symptoms, we could expect that our purely categorical disease would be hidden within a continuous symptom distribution. Can we recover such "hidden" latent classes?

Eaves et al. (1993) illustrated a model-based approach to this problem by identifying latent classes underlying the profiles of item scores over eight different items on antisocial behavior from the Rutter A-scale using children from the VTSABD (Eaves et al., 1993). Not surprisingly, a model in which all children were assumed to share a single common response profile arising from a single liability class was both implausible and fit badly. Postulating that the population was made up of a mixture of two classes of children, each with a distinct profile, fit the data much better; but postulating a mixture of three classes produced a still better fit; these are shown in Figure 3. What is striking about the profiles of the three-class solution is that they never cross. The classes are essentially ordered along a single dimension of increased liability that applies to all items (hence all the represented facets of behavior) and thus probably just reflects severity. Did this mean there were classes of severity and this was evidence for the existence of corresponding "real" categorical types of children?

The answer is "probably not." In numerical computation, a technique called Gaussian quadrature is one of the standard methods of representing a normal distribution. In this approach we replace the smoothly varying density over an infinite range of possible values by a limited number of spikes at a set of specific values, and each spike is assigned a probability weight. When used as a liability distribution, a few spikes perform like a smooth normal distribution to a surprising degree of accuracy, although for some problems a lot

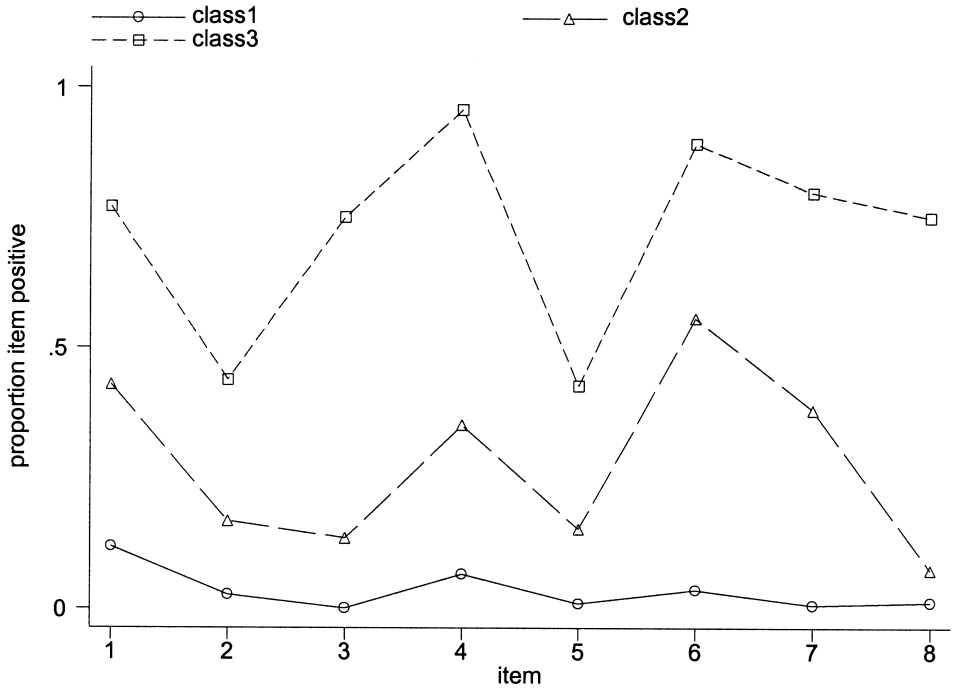

Figure 3. Item profiles for estimated latent classes from the Virginia Twin Study of Adolescent Behavioral Development (Eaves et al., 1993), Rutter A-scale items: 1 = *temper,* 2 = *steals,* 3 = *destroys,* 4 = *fights,* 5 = *not liked,* 6 = *disobeys,* 7 = *lies,* 8 = *bullies.*

of spikes may be necessary. The spikes are, mathematically speaking, identical to a set of ordered latent classes with fixed relative locations and size. If we thought of them as latent classes, then we would be assuming that the population was actually made up of classes, one corresponding to each spike. However, when we use them in Gaussian quadrature, there is no expectation that individuals "belong" to one or other of these spikes; rather, we are assuming that the population is actually normally and continuously distributed, but we are *approximating* that distribution by these spikes. Indeed, the expected score (posterior mean) of an individual on the dimension is an average (weighted by the posterior probability of belonging to each spike) of all the possible spike locations. When we plot the distribution of individual liability estimates that derive from such a model, we obtain a smooth distribution that lies between the two extreme spikes.

An intriguing extension of this approach, which corresponds exactly to ordered latent classes, is to let the location and sizes of the spikes be free parameters. Both theoretical and empirical work in statistics (e.g., Laird, 1978; Lindsay, Clogg, & Grego, 1991) has shown that the nonparametric estimator of the underlying distribution, essentially the best fitting distribution, is just such a set of discrete classes of this kind, *even when the underlying distribution is continuous.* Thus, even were children actually smoothly distributed over a continuous scale of liability, a representation in terms of a mixture of ordered categories would fit the data better than the model in which we used the correct smooth distribution! Moreover, this maximally fitting latent class representation is achieved with a remarkably small number of categories (often no more than three or four, the number depending mainly on the number of items used in the scale or profile). If one tries to fit more than this number of classes (spikes), then maximum likelihood estimation will indicate that extra class as being redundant by either assigning a zero probability weight to one of

the classes (spikes) or positioning one class in the same position as another. One consequence of all this is that results showing an ordered set of classes, such as those we have reported so far from the Eaves et al. (1993) analysis, cannot be taken as evidence for the existence of distinct "types" of children, but they may simply be the best way to statistically represent a smooth distribution.

On postulating a fourth category, Eaves et al. (1993) found that the former third category was split into two, one of children who stole and one of children who did not. On the face of it, this is more convincing evidence of distinct types of children because it is not merely a distinction in terms of severity. In practice there are reasons to doubt this, too. First, this subtyping may be reflecting not discrete categories but the fact that the children are smoothly distributed over two dimensions, the second now being a stealing dimension. (We know that the argument of the previous paragraph also applies to multidimensional data; see Davies & Pickles, 1987.) Second, the improvement in model fit from three to four latent classes, although significant, was modest. Third, classes distinguished by the complete absence of one item can result from the use of too small a sample size, and one must be very cautious about the interpretation of class profile differences under these circumstances. Kendler et al. (1998) applied similar methods to adult psychosis, identifying six classes with somewhat different symptom profiles. Caution is clearly recommended before interpreting such findings as proving a particular taxonomy or even excluding dimensional underpinnings.

Trajectories, timing, and events

Particularly in the area of antisocial behavior, researchers have become enthusiastic about a variant of this latent class approach, one in which the classes define developmental trajectories and the classification is based on a set of measures over time rather than over a single cross-section (Nagin & Land, 1993). The data required for such analyses consist of symptom or behavioral profiles obtained over a series of occasions. Previous work and theoretical considerations have led to an expectation of finding three groups: one with early-onset and persistent disorder, one adolescense limited group, and one nonantisocial group (Moffitt, 1993). In practice, however, results have been mixed; it is becoming clear that the specific groupings that are identified are dependent on the data that are chosen or available for analysis. For example, if the window of observation is extended beyond early adulthood, then additional desistance and late-onset classes can be identified. The approach does provide a sound framework within which to compare models that impose alternative restrictions on the number or form of classes, but as in all such tests there is an assumption that the overall class of models being considered includes one that is correct. Moreover, the approach does not explicitly consider dimensional alternatives that could yield equally well fitting models, nor that some of the models may be close approximations to a nonparametric maximum likelihood representation, one that we have seen is quite neutral on the category or dimension question.

As we have just seen, among the distinguishing features of trajectories are differences in age of onset and desistance. Methods of analysis specific to the timing of events have been developed, in the form of survival and event history models (Allison, 1984; Cox, 1972), and differences in age of onset have been used to distinguish qualitatively different forms of pathology, as when early onset dementia was linked to APO-E4 (reviewed in Ritchie & Dupuy, 1999). Event history methods typically presuppose that a categorical phenomenon exists with an onset occurring in a sharply localized period of time. However, much psychopathology is not of this kind. There may be early-onset prodromal symptoms, or symptom severity may increment progressively over time. Whether we can consider a developmental onset as an event may depend on our study design. For example, at age 6, very few girls are biologically competent to become pregnant, but by age 16 most are. Thus, in a comparison of 6- and 16-year-olds, pubertal status has a strongly bimodal distribution and some sort of transitional "event" (i.e., puberty) has apparently occurred. However, if one were studying just 11-year-

old girls, a wide range of pubertal statuses would be observed and this condition might best be regarded as being dimensionally distributed. When viewed over the longer period, pubertal development could be treated as an event and its timing analyzed using survival analysis methods. Even when restricted to the teenage years, some aspects of pubertal development such as menarche remain more event-like, but others, such as breast development are often better approached using growth curves. Pickles et al. (1996, 1998) contrasted these two types of event onset as "hard" and "soft"; they showed that, although they seemed to be under the control of a common set of genes, when gathered by means of retrospective report, these two types of events suffered different forms of measurement error (Huttenlocher, Hedges, & Prochaska, 1988). Hard events suffered heteroscedastic measurement error (random error that increases with time since the event), but soft events were prone to telescoping (systematic bias in which the reported time is moved toward the time at which the report was elicited). Overall, therefore, interest in timing and age of onset favors the presumption of a categorical phenomenon but does not exclude underlying dimensional variation. Nonetheless, the researcher may need to remain sensitive to the possibly differing measurement issues that arise with hard and soft events.

Nonlinearity

If the distribution of the indicators of the pathology alone cannot help, perhaps we should expand the scope of potential evidence by looking for some sort of validator. What can the relationship of pathology to some other variable tell us about the nature of that pathology? Simple linear logistic regressions provide estimates of effects that average across the whole sample ranges of the predictor variables, whereas ordinary correlations and linear regressions estimate average relationships across the ranges of both the predictors and the outcome variable. It is also the case that a computer statistical package will fit a linear model of this sort if asked of it, even if any relationship is shared by only part of the range of the predictor variable. Without deliberate effort, the routine use of these standard statistical techniques rarely bring such discontinuities to the researcher's attention, but they may be important indicators of the presence of categorical states.

An essential feature of the categorical disease entity model (at least as it is implemented in the *DSM-IV*) is that the categories should be associated with functional impairment (Wakefield, 1992, 1997). Pickles et al. (2001), again using the VTSABD, examined how symptom related impairment increased with the number of symptoms. For conduct disorder, oppositional defiant disorder, and depression, there was a smooth linear relationship. There was no evidence of any discontinuity or jump in impairment associated with the *DSM-III-R* symptom cutoffs, or any other plausible symptom cutoff. Impairment increased with severity, but disorder per se added nothing more. Had it been found that symptoms at or above the threshold were associated with a more marked increase in impairment than those below the threshold, then could this have been taken as evidence of a discrete pathology? Again, the answer is "not necessarily." As with inference from the shape of the symptom distribution itself, consideration still has to be given to whether a transformation of the impairment scale was appropriate and whether applying such a transformation would have eliminated the evidence for nonlinearity.

Risk to other relatives has also been used as a potential validator of a category. Using the Twins Early Development Study, Dale et al. (1998) applied the DeFries and Fulker (1985) regression method to assess differential heritability of language problems: whether heritability increased or decreased as the severity of problems experienced by the proband increased. In a general population sample of UK twins, they found that language skills at age 2 were predominantly influenced by environmental factors but that genetic factors appeared to be of greater importance for the most severely delayed. This provided some support for the distinctiveness of severe language problems. However, here, the results of the DeFries–Fulker method can also be quite sensitive to

the choice of transformation applied to the phenotypic measure.

Maximizing within category association

The previous section hints that the kind of evidence that could be persuasive of the existence of categories involves showing that different categories possess distinctive patterns of association (Meehl, 1992). Meehl (Meehl, 1995; Waller & Meehl, 1998) formalizes this approach into the maximum covariance (MAXCOV) criterion. This requires that we consider at least three variables, say x, y, and z. A priori considerations suggest that one of these, say x, is thought to be an indicator of a binary categorical variable rather than a measure of some underlying dimensional score. If correct, then we would expect those in one category to have lower values and those in the other category to have higher values of x. Moreover, on the supposition that categories are internally homogeneous, we would expect that if we considered each category separately, there would be a rather modest covariance between y and z, whereas if we took a mix of subjects from both categories, we would observe a more substantial y–z covariance. For example, x could be a symptom score and y and z could be two measures of risk, say parental neglect and problems with peers. Within the disorder group we might expect both risk factors to be raised, whereas in the normal group we would expect both to be low, resulting in low within group covariances. Only when the groups are mixed do we see that the two risk factors covary together.

Meehl formalizes this into a criterion. The y–z covariance is calculated for sets of subjects defined by a range or window of values of x. This is repeated several times, moving the window across the distribution of x. Plotted (with suitable smoothing) against the mid-value of each window, the y–z covariance will oscillate randomly if there is no category underlying x, but it will increase to a maximum and then decrease if a category underlies x, the MAXCOV being that obtained from the window position in which subjects from each category are equally represented. The plot thus allows both the categorical nature and the prevalence of the categories to be identified.

Do we see such patterns in real samples? Waller, Putnam, and Carlson (1996) applied the MAXCOV criterion (and other related methods) to data from a case-normal control study of dissociative identity disorder and found strong evidence for a categorically distinct pattern of association among a subset of the symptoms of the Dissociative Experiences Scale (Bernstein–Carlson & Putnam, 1986). In this application one cannot help feeling, however, that the selection of subjects from the extremes that is a consequence of the case-normal control design is likely to favor taxon hunters. However, perhaps more importantly, we should question the underlying model. First, the expectation of no within category association is an assumption and there would seem to be many circumstances where our scientific theory might suggest categories that are distinguished, not by having a common lack of association between y and z, but by them having distinctively different patterns of association, say, being positively associated in one category and not, or even negatively, associated in the other. For example, in a study of the symptomatology shown by relatives of autistic probands, Pickles et al. (2001) found that the proportion of affected relatives appeared to increase with the severity of autism shown by the proband, but only among probands with "useful speech." This raised the possibility that speech was a marker for etiologically distinct forms of autism. Second, the categoric nature of x clearly depends upon the choice of y and z. What do we conclude if we apply the MAXCOV approach with variables x, y, and z and find evidence for categories underlying x, but then apply the same approach to x, u, and v and find evidence for x being continuous?

Dimensional or Categorical Nature Versus Dimensional and Categorical Properties: Depression in the Great Smoky Mountains Study

The question above is resolved when we recognize that we never directly observe our objects of interest but instead observe their properties. Different properties do not need to conform to a single conceptual model of the pathology.

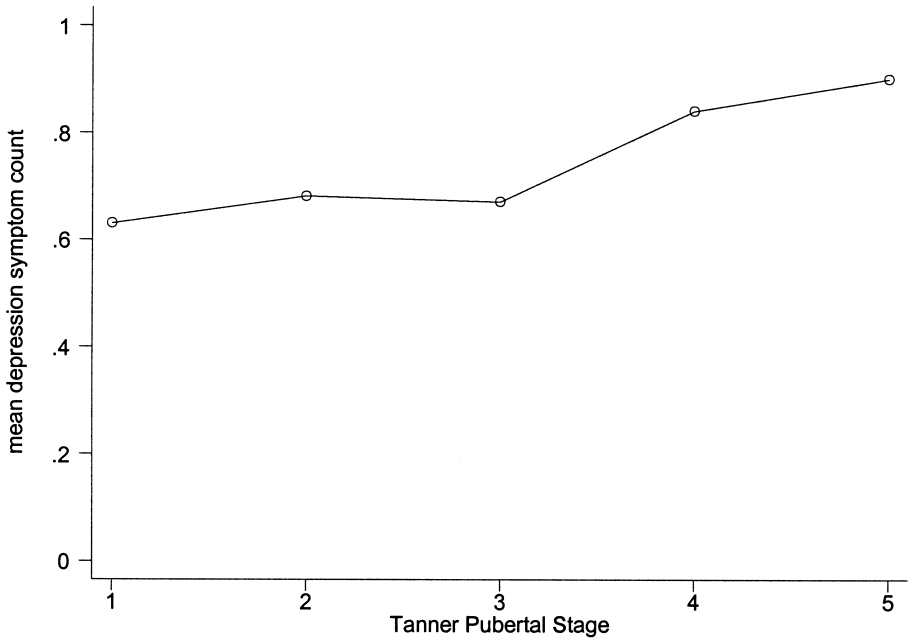

Figure 4. The mean *DSM-IV* depression symptom count by pubertal stage for girls from the Great Smoky Mountains Study.

We illustrate this point with data from the Great Smoky Mountains Study. We examine first the relationship between puberty and depression in girls. We began this line of research in an attempt to clarify the timing and causes of the emergence of the female preponderance of depression observed in adults, and have concluded that the most potent factor appears to be increasing levels of sex steroid hormones acting through a mechanism unrelated to effects on secondary sex characteristics (Angold, Costello, & Worthman, 1998, 1999; Angold & Worthman, 1993). Here we present some additional analyses of this topic designed to illustrate some of the analytical issues we have discussed thus far.

First, we note that *DSM-IV* depression and dysthymia symptom counts were "continuously" (but certainly not normally) distributed, with no obvious points of rarity or bulges in the distribution. As shown in Figure 4, the count also increased with Tanner stage, but this increase was not linear on the Tanner scale. The counts were very similar at Stages 1, 2, or 3 but higher at stages 4 and 5. If we ignore this and simply fit a linear regression model, we find no significant relationship between the Tanner stage and the number of depressive symptoms ($p = .3$).

However, if one divides the Tanner stage into two categories formed by grouping stage 3 or below versus stage 4 and above, the significance of the effect *increases* ($p = .07$). Taking this analysis a step further, we can examine the outcome variable (depressive symptom count) in a similar way. However, it turns out that the relationship between pubertal stage and symptom count is *not uniform across the range of symptom counts*. Figure 5 shows that there is no relationship between pubertal stage and the probability of having just one symptom. All of the effect of puberty on symptoms is carried by those with two or more symptoms, odds ratio (OR) = 1.7; 95% confidence interval (CI) 1.2–2.5, $p = .008$. Thus, it turns out that we are dealing with a relationship that is most parsimoniously described as categorical, even though both the predictor and outcome variables were measured on dimensional scales.

It is well known that if a genuinely dimensional relationship exists between two phenomena, then modeling it as a categorical relationship is wasteful of information (and

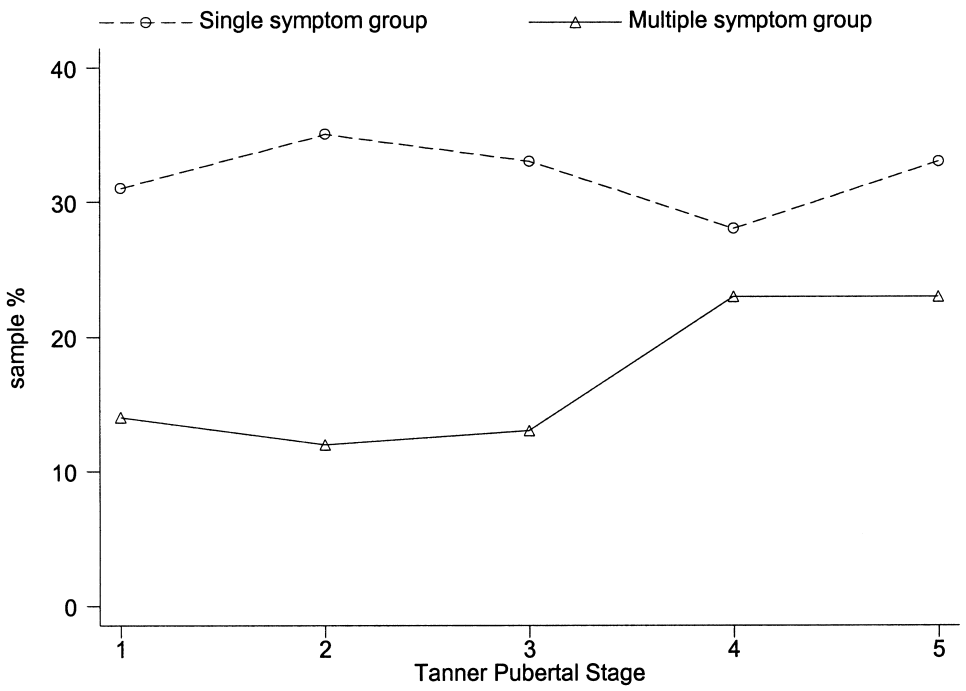

Figure 5. Sample proportions with one or more *DSM-IV* symptoms of depression by pubertal stage for girls, from the Great Smoky Mountains Study.

therefore power). Thus, if a continuous relationship exists across the range of two variables, then performing a median split on each and testing the relationship using a 2×2 contingency table will underestimate the size of the relationship. However, it seems to be less appreciated that the same is true if one imposes a continuous metric on a categorical relationship.

We next turn to examine the relationship of the count of depressive symptoms to a risk index defined by a summary scale of 26 social, family, and life event risk factors for psychopathology. Here we observe a quite different pattern. In this case, we see, in Figure 6, what looks like a genuinely dimensional relationship. No cut-point on the predictor risk factor scale or the depression scale produced a larger test statistic for the association between these two variables than the linear association of the two (means ratio from Poisson regression = 1.14; 95% CI = 1.10–1.19; $p < .0001$). What we see is that the "same psychopathology" may apparently behave both categorically and dimensionally in relation to different risk factors.

Conclusions

It is perhaps unsurprising that psychopathologists continue to debate the issues surrounding the use of categorical and dimensional perspectives. Their workplaces and research centers are dominated by medical doctors, who draw on a long tradition of discrete medical diagnoses and therapeutic interventions, and by psychologists, who are part of a long tradition of population measurement and continuous scale score construction. It is hard to emphasize just how profound are the consequences of these contrasting perspectives for the cultures of these two disciplines.

There is a strong desire for our chosen taxonomy to be "right." Meehl approvingly quotes the aphorism from Plato, that we should "cut nature at its joints" and points to the fact that we have gophers and chipmunks but we do not have "gophmunks." We cannot add, mix,

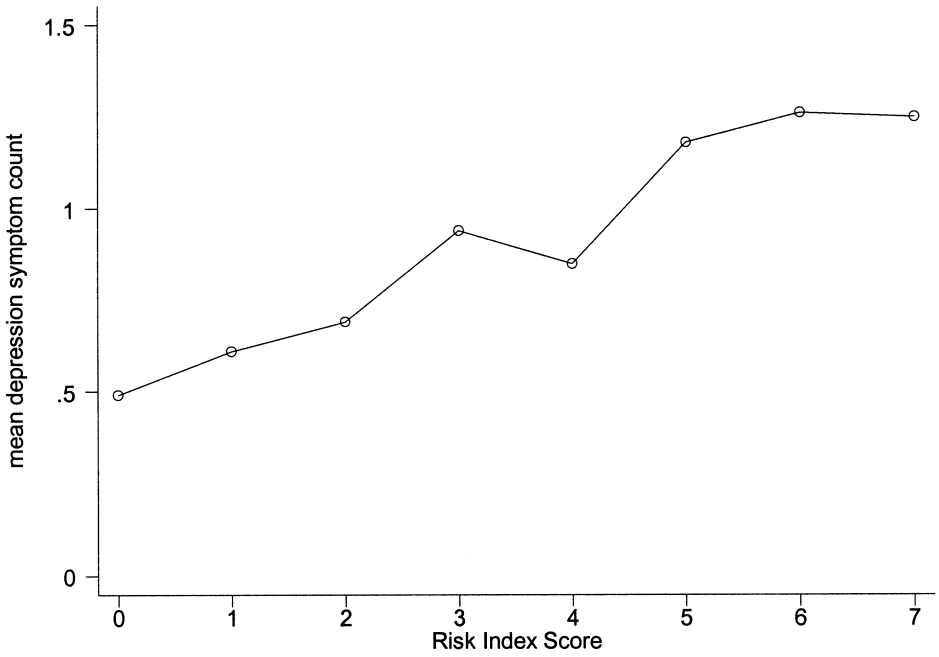

Figure 6. The mean *DSM-IV* depression symptom count by risk index score for girls from the Great Smoky Mountains Study.

and average these qualitatively distinct species as we could were the distinctions between them quantitative. While it should be emphasized that Meehl does not argue that there are not also important dimensions of variation, there is nonetheless the sense that, once a qualitative facet has been found, the objects or phenomena are therefore categorical in nature. Before drawing such a conclusion, it is worth noting that species boundaries are by no means so distinct and in the process of evolution gophmunks might very well appear! After all, the whole notion of evolution first required acceptance of the mutability of species. More importantly, we believe that the species analogy places the focus of attention on some abstract state of nature of the object of study, rather than on the properties that it exhibits. It is through these properties that we largely define objects of study, the properties are many and it is they that are also commonly of direct scientific and clinical interest. Crucially, our argument is that these varied properties need not be consistent with a single state of nature. Rather than the analogy with species, we prefer an analogy with high school theories of the duality of light. When it comes to understanding refraction, a conception of light as behaving as a continuous wave turns out to be helpful, whereas particulate behavior is more helpful conceptually for understanding reflection and energy transfer. In the same way, to understand depression, we find circumstances where it behaves as a dimension and circumstances where it behaves as a more discrete phenomenon. Thus, while we accept the absence of gophmunks, we can envisage purposes, such as making Brunswick Stew, where we are merely concerned with weight, and the difference between a small gopher and a large chipmunk counts for nothing (except among those who believe that only squirrel should appear in Brunswick Stew). The important issue is not whether depression is categorical or dimensional in any general sense, but how its relationships with etiological, outcome, and other factors are manifested. If we are to answer those questions properly, then we need to keep an open mind about the shapes of the associational curves. In other words, we

need to adopt a truly empirical approach that is unblinkered by either categorical or dimensional prejudices.

There are a number of consequences of holding this view. First, we should strive against the metatheoretical implications of a choice that results in us unconsciously selecting whole superstructures of thought and practice when we choose to use categories or dimensions. Second, although there may well be circumstances in which one perspective is clearly more parsimonious, these are probably comparatively rare and proving the empirical advantage of one or the other is nontrivial, and may not often be worth the trouble. Thus, we will typically find that we need to entertain both conceptions simultaneously, exploring analyses and elaborating theories that are not exclusively in one tradition or the other. Third, we should exploit the advantages of each perspective as and when it is opportune. Those advantages could relate to the ease of operationalizing a theory, ease of measurement, ease of analysis, or merely ease of communicating the results. However, in so doing, we need to remain cautious and to avoid both over interpretation and implying that alternative conceptions have been empirically excluded, when they have not. Fourth, as methodologists we should also be working to develop a more fluid set of tools to assist in analysis and communication in this bilingual world.

References

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs, 80,* 1–37.

Achenbach, T. M. (1985). *Assessment and taxonomy of child and adolescent psychopathology*. Beverly Hills, CA: Sage.

Achenbach, T. M. (1991a). *Child Behavior Checklist: Teacher's Report Form*. Burlington, VT: University of Vermont, Center for Children, Youth, and Families.

Achenbach, T. M. (1991b). The derivation of taxonomic constructs: A necessary stage in the development of developmental psychopathology. In D. Cicchetti & S. Toth (Eds.), *Rochester symposium on developmental psychopathology: Vol. 3. Models and integrations.* Rochester, NY: University of Rochester Press.

Achenbach, T. M. (1991c). *Manual for the Child Behavior Checklist 4–18 and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2–3*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Conners, C. K., Quay, H. C., Verhulst, F. C., & Howell, C. T. (1989). Replication of empirically derived syndromes as a basis for taxonomy of child/adolescent psychopathology. *Journal of Abnormal Child Psychology, 17,* 299–323.

Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Review, 85,* 1275–1301.

Achenbach, T. M., & Edelbrock, C. (1991). *Child Behavior Checklist: Youth Self-Report*. Burlington, VT: University of Vermont, Center for Children, Youth, and Families.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA Preschool Forms and Profiles: An integrated system of multi-informant assessment*. Burlington, VT: University of Vermont, Department of Psychiatry.

Allison, P. D. (1984) *Event history analysis. Regression for longitudinal event data*. Newbury Park, CA: Sage.

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: American Psychiatric Press.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: American Psychiatric Press.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Press.

Angold, A., & Costello, E. J. (1996). Toward establishing an empirical basis for the diagnosis of Oppositional Defiant disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 1205–1212.

Angold, A., Costello, E. J., & Worthman, C. M. (1998). Puberty and depression: The roles of age, pubertal status, and pubertal timing. *Psychological Medicine, 28,* 51–61.

Angold, A., Costello, E. J., & Worthman, C. M. (1999). Pubertal changes in hormone levels and depression in girls. *Psychological Medicine, 29,* 1043–1053.

Angold, A., & Fisher, P. W. (1999). Interviewer-based interviews. In D. Shaffer, C. Lucas, & J. Richters (Eds.), *Diagnostic assessment in child and adolescent psychopathology* (pp. 34–64). New York: Guilford Press.

Angold, A., & Worthman, C. W. (1993). Puberty onset of gender differences in rates of depression: A developmental, epidemiologic and neuroendocrine perspective. *Journal of Affective Disorders, 29,* 145–158.

Berger, M., Yule, W., & Rutter, M. (1975). Attainment and adjustment in two geographical areas II. The prevalence of specific reading retardation. *British Journal of Psychiatry, 126,* 510–519.

Bergman, L. R. (2001) A person approach in research on adolescence: Some methodological challenges. *Journal of Adolescent Research 16,* 28–53.

Bernstein–Carlson, E. B., & Putnam, F.W. (1986). Development, reliability, and validity of a dissociation scale. *Journal of Nervous and Mental Disease, 174,* 727–735.

Blashfield, R. K. (1982). Feighner et al., invisible colleges, and the Matthew effect. *Schizophrenia Bulletin, 8,* 1–12.

Cairns, R. B., & Green, J. A. (1979). How to assess personality and social patterns: Ratings or observations? In R. B. Cairns (Ed.), The analysis of social interaction: Methods, issues and illustrations. Hillsdale, NJ: Erlbaum.

Conners, C. K. (1997). *Conners' Rating Scales Revised: Instruments for use with children and adolescents*. North Tonawanda, NY: Multi-Health Systems.

Costello, E. J., Angold, A., Burns, B. J., Stangl, D. K., Tweed, D. L., Erkanli, A., & Worthman, C. M. (1996). The Great Smoky Mountains Study of Youth. Goals, design, methods, and the prevalence of *DSM-III-R* disorders. *Archives of General Psychiatry, 53,* 1129–1136.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B, 74,* 187–200.

Crijnen, A., Achenbach, T. M., & Verhulst, F. C. (1997). Comparisons of problems reported by parents of children in 12 cultures: Total problems, externalizing, and internalizing. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1269–1277.

Cullen, K. J., & Boundy, C. A. P. (1966). The prevalence of behavior disorders in the children of 1000 Western Australian families. *Medical Journal of Australia, 2,* 805–808.

Cummings, J. D. (1944). The incidence of emotional symptoms in school children. *British Journal of Educational Psychology, 14,* 151–161.

Dale, P. S., Simonoff, E., Bishop, D. V. M., Eley, T. C., Oliver, B., Price, P.S., Purcell, S., Stevenson, J., & Plomin, R. (1998). Genetic involvement in language delay in 2 year-old children. *Nature Neuroscience, 1,* 324–328.

Davies, R. B., & Pickles, A. R. (1987). A joint trip-timing/store-type choice model for grocery shopping including feedback, inventory effects and non-parametric control for omitted variables. *Transportation Research, 21,* 345–361.

DeFries, J. C., & Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behavior Genetics, 15,* 467–473.

Eaves, L. J., Silberg, J. L., Hewitt, J. K., Rutter, M., Meyer, M., Neale, M., & Pickles, A. (1993). Analyzing twin resemblance in multisymptom data: Genetic application of latent class models for symptoms of conduct disorder in juvenile boys. *Behavior Genetics, 23,* 5–19.

Gould, M. S., Wunsch–Hitzig, R., & Dohrenwend, B. (1981). Estimating the prevalence of child psychopathology. *Journal of the American Academy of Child Psychiatry, 20,* 462–476.

Graham, P., & Rutter, M. (1973). Psychiatric disorders in the young adolescent: A follow-up study. *Proceedings of the Royal Society of Medicine, 6,* 1226–1229.

Griffiths, W. (1952). *Behavior difficulties of children as perceived and judged by parents, teachers, and children themselves*. Minneapolis, MN: The University of Minnesota Press.

Haggerty, M. E. (1925). The incidence of undesirable behavior in public school children. *Journal of Educational Research, 12,* 102–122.

Hill, J., Pickles, A., Burnside, E., Byatt, M., Rollinson, L., Davis, R., & Harvey, K. (2001). Child sexual abuse, poor parental care and adult depression: Evidence for different mechanisms. *British Journal of Psychiatry, 179,* 110–115.

Huttenlocher, J., Hedges, L., & Prochaska, V. (1988). Hierarchical organisation in ordered domains: Estimating dates of events. *Psychological Review, 95,* 471–484.

Kendall, M. (1991). The major functional psychoses: Are they independent entities or part of a continuum? Philosophical and conceptual issues underlying the debate. In A. Kerr & H. McLelland (Eds.), *Concepts of mental disorder: A continuing debate* (pp. 1–16). London: Gaskell.

Kendler, K., Karkowski, L., & Walsh, D. (1998). The structure of psychosis: Latent class analysis of probands from the Roscommon Family Study. *Archives of General Psychiatry 55,* 492–499.

Kohlberg, L. (1976). Moral stages and moralization. In T. Lickona (Ed.), *Moral development and behavior* (pp. 31–53). New York: Holt Rinehart & Winston.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods, 4,* 257–271.

Laird, N. M. (1978). Non-parametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73,* 805–811.

Lapouse, R. (1966). The epidemiology of behavior disorders in children. *American Journal of Disfunctional Children, 111,* 594–599.

Lapouse, R. L., & Monk, M. A. (1958). An epidemiologic study of behavior characteristics in children. *American Journal of Public Health, 48,* 1134–1144.

Lapouse, R. L., & Monk, M. A. (1964). Behavior deviations in a representative sample of children: Variation by sex, age, race, social class and family size. *American Journal of Orthopsychiatry, 34,* 436–446.

Lindsay, B. G., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96–107.

Long, A. (1941). Parents' reports of undesirable behavior in children. *Child Development, 12,* 43–62.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

Macready, G. B., & Dayton, C. M. (1994). Latent class models for longitudinal assessment of trait acquisition. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis* (pp. 245–273). Thousand Oaks, CA: Sage.

Maddala, G. (1983). *Limited dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Magnusson, D. (1988a). Methodological and research strategical considerations of an interactional perspective. In D. Magnusson (Ed.), *Paths through life: Vol. 1. Individual development from an interactional perspective: A longitudinal study* (pp. 62–82). Hillsdale, NJ: Erlbaum.

Magnusson, D. (1988b). A theoretical framework: The interactional perspective. In D. Magnusson (Ed.), *Paths through life: Vol. 1. Individual development from an interactional perspective: A longitudinal study* (pp. 15–43). Hillsdale, NJ: Erlbaum.

McFie, B. S. (1934). Behavior and personality difficulties in school children. *British Journal of Educational Psychology, 4,* 34.

Meehl, P. E. (1992). Factors, and taxa, traits and types,

difference of degree and differences of kind. *Journal of Personality, 60,* 117–174.

Meehl, P. E. (1995). Bootstrap taxometrics: Solving the classification problem in psychpathology. *American Psychologist, 50,* 266–275.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent-offending: A complementary pair of developmental theories. In T. P. Thornberry (Ed.), *Developmental theories of crime and delinquency* (pp. 11–55). New Brunswick, NJ: Transaction.

Muthen, B., & Muthen, L. (2000). Integrating person-centred and variable-centred analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24,* 882–891.

Muthén, L. K., & Muthén, B. (2001). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers and population heterogeneity: Specification and estimation of a non-parametric mixed Poisson model. *Criminology, 31,* 327–362.

Offord, D. R., Kraemer, H. C., Kazdin, A. E., Jensen, P. S., & Harrington, R. (1998). Lowering the burden of suffering from child-psychiatric disorder: Trade-offs among clinical, targeted, and universal interventions. *Journal of the American Academy of Child and Adolescent Psychiatry, 37,* 686–694.

Olson, W. C. (1930). *Problem tendencies in children*. Minneapolis, MN: University of Minnesota Press.

Piaget, J. (1932). *The moral judgment of the child*. New York: Harcourt, Brace & Company.

Pickles A. R. (1989). Statistical modelling of longitudinal data. In M. Rutter (Ed.), *The power of longitudinal data: Studies of risk and protective factors for psychosocial disorder* (pp. 62–76). Cambridge: Cambridge University Press.

Pickles, A. (1993). Stages, precursors and causes in development. In D. F. Hay & A. Angold (Eds.), *Precursors and causes in development and psychopathology* (pp. 23–50). Chichester, UK: Wiley.

Pickles, A., & Clayton, D. G. (2002). *Log-odds estimates for mis-measured exposures using normal theory random effects models.* Unpublished manuscript, University of Manchester.

Pickles, A., Pickering, K., Simonoff, E., Meyer, J., Silberg, J. & Maes, H. (1998). Genetic "clocks" and "soft" events: A twin model for pubertal development and other recalled sequences of developmental milestones. *Behavior Genetics, 28,* 243–253.

Pickles, A., Pickering, K., & Taylor, C. (1996). Reconciling recalled dates of developmental milestones, events and transitions: A mixed GLM with random mean and variance functions. *Journal of the Royal Statistical Society* Series *A, 159,* 225–234.

Pickles, A., Rowe, R., Simonoff, E., Foley, D., Rutter, M. N., & Silberg, J. (2001). Child psychiatric symptoms and psycho-social impairment: Relationship and prognostic significance. *British Journal of Psychiatry, 179,* 230–235.

Quay, H. C. (1977). Measuring dimensions of deviant behavior: The Behavior Problem Checklist. *Journal of Abnormal Child Psychology, 5,* 277–287.

Rabe–Hesketh, S., Pickles, A., & Skrondal, A. (2002). *Correcting for covariate measurement error in logistic regression using a nonparametric maximum likelihood estimation.* Manuscript submitted for publication.

Ritchie, K., & Dupuy, A. M. (1999) . The current status of APO E4 as a risk factor for Alzheimer's disease: An epidemiological perspective. *International Journal of Geriatric Psychiatry, 14,* 695–700.

Rutter, M. (1965). Classification and categorization in child psychiatry. *Journal of Child Psychology and Psychiatry, 6,* 71–83.

Rutter, M. (1976). Research report: Isle of Wight studies 1964–1974. *Psychological Medicine, 6,* 313–332.

Rutter, M., Graham, P., Chadwick, O. F. D., & Yule, W. (1976). Adolescent turmoil: Fact or fiction? *Journal of Child Psychology and Psychiatry, 17,* 35–56.

Rutter, M., Lebovici, S., Eisenberg, L., Sneznevskij, A. V., Sadoun, R., Brooke, E., & Lin, T. Y. (1969). A tri-axial classification of mental disorders in childhood: An international study. *Journal of Child Psychology and Psychiatry, 10,* 41–61.

Rutter, M., & Shaffer, D. (1980). *DSM-III*: A step forward or back in terms of the classification of child psychiatric disorders? *Journal of the American Academy of Child Psychiatry, 19,* 371–394.

Rutter, M., Tizard, J., & Whitmore, K. (1970). *Education, health, and behaviour*. London: Longman.

Rutter, M., Yule, B., Morton, J., & Bagley, C. (1975). Children of West Indian immigrants III: Home circumstances and family patterns. *Journal of Child Psychology and Psychiatry, 16,* 105–123.

Rutter, M., Yule, B., Quinton, D., Rowlands, O., Yule, W., & Berger, M. (1974). Attainment and adjustment in two geographical areas III. Some factors accounting for area differences. *British Journal of Psychiatry, 125,* 520–533.

Rutter, M., Yule, W., Berger, M., Yule, B., Morton, J., & Bagley, C. (1974). Children of West Indian immigrants: I. Rates of behavioural deviance and of psychiatric disorder. *Journal of Child Psychology and Psychiatry, 15,* 241–262.

Rutter, M. L., Shaffer, D., & Sturge, C. (1979). *A guide to a multi-axial classification scheme for psychiatric disorders in childhood and adolescence*. London: Frowde & Co. Ltd.

Schumacker, R. E., & Marcoulides, G. A. (1998). *Interaction and nonlinear effects in structural equation modeling*. Mahwah, NJ: Erlbaum.

Shaffer, D., Fisher, P. W., & Lucas, C. P. (1999). Respondent-based interviews. In D. Shaffer, C. P. Lucas, & J. E. Richters (Eds.), *Diagnostic assessment in child and adolescent psychopathology* (pp. 3–33). New York: Guilford Press.

Sonuga–Barke, E. J. S. (1998). Categorical models of childhood disorder: A conceptual and empirical analysis. *Journal of Child Psychology and Psychiatry, 39,* 115–133.

Srole, L., Langner, T. S., Mitchell, S. T., Opler, M. K., & Rennie, T. A. C. (1962). *Mental health in the metropolis: The Midtown Study*. New York: McGraw–Hill.

Taylor, E. (1994). Similarities and differences in *DSM-IV* and *ICD-10* diagnostic criteria. *Child and Adolescent Psychiatric Clinics of North America, 3,* 209–226.

Van den Oord, E., Pickles, A., & Waldman, I. (2003). Normal variation and abnormality: An empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry, 44,* 180–192.

Verhulst, F. C., & Achenbach, T. M. (1995). Empirically based assessment and taxonomy of psychopathology: Cross-cultural applications. A Review. *European Child and Adolescent Psychiatry, 4,* 61–76.

von Eye, A. (1990). *Introduction to configural frequency*

analysis: The search for types and anti-types in cross-classifications. Cambridge: Cambridge University Press.

Wakefield, J. C. (1992). Disorder as harmful dysfunction: A conceptual critique of *DSM-III-R*'s definition of mental disorder. *Psychological Review, 99,* 232–247.

Wakefield, J. C. (1997). When is development disordered? Developmental psychopathology and the harmful dysfunction analysis of mental disorders. *Development and Psychopathology, 9,* 269–290.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures. Distinguishing types from continua*. Thousand Oaks, CA: Sage.

Waller, N. G., Putnam, F. W., & Carlson, E. (1996). Types of dissociation and dissociative types: A taxonometric analysis of the Dissociative Experiences Scale. *Psychological Methods, 3,* 300–321.

Wilson, M. (1993). *DSM-II* and the transformation of American psychiatry: A history. *American Journal of Psychiatry, 150,* 399–410.

World Health Organization (1993). *ICD-10 Classification of mental and behavioural disorders: Diagnostic criteria for research*. Geneva: World Health Organization.

World Health Organization (1994). *ICD-10 Classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Wickman, E. K. (1928). *Children's behavior and teachers' attitudes*. New York: Commonwealth Fund.

Young–Masten, I. (1938). Behavior problems of elementary school children: A descriptive and comparative study. *Genetic Psychology Monographs, 20,* 123–180.

Yourman, J. (1932). Children identified by their teachers as problems. *Journal of Educational Sociology, 5,* 334–343.

Yule, W., Berger, M., Rutter, M., & Yule, B. (1975). Children of West Indian immigrants: II. Intellectual performance and reading attainment. *Journal of Child Psychology and Psychiatry, 16,* 1–17.