Psychological Review
2005, Vol. 112, No. 1, 129–158

0033-295X/05/$12.00 DOI: 10.1037/0033-295X.112.1.129

A Conceptual and Psychometric Framework for Distinguishing Categories and Dimensions

Paul De Boeck
K. U. Leuven

Mark Wilson
University of California, Berkeley

G. Scott Acton
Rochester Institute of Technology

An important, sometimes controversial feature of all psychological phenomena is whether they are categorical or dimensional. A conceptual and psychometric framework is described for distinguishing whether the latent structure behind manifest categories (e.g., psychiatric diagnoses, attitude groups, or stages of development) is category-like or dimension-like. Being dimension-like requires (a) within-category heterogeneity and (b) between-category quantitative differences. Being category-like requires (a) within-category homogeneity and (b) between-category qualitative differences. The relation between this classification and abrupt versus smooth differences is discussed. Hybrid structures are possible. Being category-like is itself a matter of degree; the authors offer a formalized framework to determine this degree. Empirical applications to personality disorders, attitudes toward capital punishment, and stages of cognitive development illustrate the approach.

In this article we describe a conceptual and psychometric scheme for distinguishing the categorical versus dimensional nature of psychological variables. By *psychological variables* we mean variables used to distinguish between entities in some psychological respect. These entities are commonly persons, but they can be also situations, tasks, test items, and so on. The scheme arose out of frustrations with instances of psychological research that had either assumed or "proved" that variables were of one kind or the other without examining the philosophical or empirical basis for doing so and without providing an overarching framework in which either was genuinely possible. In this article, we provide such an overarching framework, called the *dimension/category framework* (Dimcat), and provide empirical illustrations of its use.

A preliminary distinction in determining whether variables are category-like or dimension-like is the distinction between *manifest variables* and *latent variables.* Too often these two kinds of variables are confused, which can lead to inappropriate conclusions. Specifically, researchers may confuse manifest categories or dimensions, which are artifacts of the measurement approach, with latent categories or dimensions, which are typically the underlying psychological phenomena of interest.

The issue under consideration here is whether the latent nature of manifest variables is category-like or dimension-like. One assumption might be that the nature of the latent and manifest variables match. As discussed below, however, manifest dimensions can be turned into manifest categories (e.g., in segmentation into groups), and manifest categories can be turned into manifest dimensions (e.g., in sum scores on a test). Thus, the relations between different kinds of manifest variables and between different kinds of manifest and latent variables are not so simple as they might at first appear. Consequently, a conceptual and methodological framework that encompasses all of these possibilities is needed.

Manifest dimensions (or *manifest continua*) are common in psychological research, although their dimensional nature may be only a convenient fiction. For example, raw scores on a test (e.g., number of correct responses) are ordered manifest categories, yet they are commonly seen as approximating a manifest dimension. Items on a test are examples of *indicators* in the same way that symptoms in a diagnostic system are indicators, although these different kinds of indicators are typically put to very different uses. Whereas items are typically summed to produce a manifest dimension, symptoms are typically summed to produce a manifest category (a diagnosis). To complicate matters, a manifest dimension based on item sums may also be segmented (e.g., using a median split) to produce a manifest category, or the sum of symptoms may be used as an indicator of the extent to which patients show a syndrome. It should be apparent from this discussion that manifest

Paul De Boeck, Department of Psychology, K. U. Leuven, Leuven, Belgium; Mark Wilson, Graduate School of Education, University of California, Berkeley; G. Scott Acton, Department of Psychology, Rochester Institute of Technology.

Work on this article was supported in part by grants to Paul De Boeck from K. U. Leuven (GOA-00/2) and the Belgian National Fund for Scientific Research, by a grant to Mark Wilson from the Spencer Foundation, and by grants to G. Scott Acton from the National Institute on Drug Abuse (F32-DA14739 and P50-DA09253) and the California Tobacco-Related Disease Research Program (10FT-0248).

We thank Christine Maesschalck and Kim Pottie for gathering data, and Katalin Balazs, Dorien Dossche, Istvan Hidegkuti, Michel Meulders, Frank Rijmen, Dirk Smits, Francis Tuerlinckx, and Timothy Verbeemen for help in various respects. Paul De Boeck and Mark Wilson contributed equally to this article.

Correspondence concerning this article should be addressed to Paul De Boeck, Department of Psychology, K. U. Leuven, Tiensestraat 102, B-3000, Leuven, Belgium. E-mail: paul.deboeck@psy.kuleuven.ac.be

categories and manifest dimensions can be functionally interchangeable and thus arbitrary.

Latent dimensions are quantitative variables with values that depend on the person and that in one way or another contribute to the observations, either (a) directly or (b) indirectly via the effect the quantitative variable has on the probability of the responses. For a discussion of the epistemological status of latent variables, see Borsboom, Mellenbergh, and van Heerden's (2003) article. Latent dimensions are invoked as underlying quantities that determine data or functions thereof, such as the sum score. For example, in classical test theory, a true score (latent dimension) is believed to be at the basis of the sum score of a test (manifest dimension), except for distortions due to the so-called error term. Latent dimensions are implicit whenever concepts like internal-consistency reliability are used—that is, in virtually all tests of psychological phenomena. The underlying variables in factor analysis models, structural equation modeling, and item response theory (IRT) are not manifest but latent dimensions.

Manifest categories are also common in psychological research, as independent or dependent variables. Regardless of whether the categorical variables are independent or dependent variables, they are often (but not always) rooted in, derived from, based on, or linked to some manifest or tacit indicators from the same domain. Indicators need to be either directly or indirectly observed for one to derive a manifest category from them.

A manifest category is commonly derived from indicators through either segmentation or expert judgment. *Segmentation* means that one indicator or a composite of indicators (e.g., a sum score on a test) is segmented into different manifest categories. Some segments may be omitted, as in the method of extreme groups, in which the middle segment is omitted. *Expert judgment* means that an expert attributes manifest categories on the implicit or explicit basis of knowledge regarding the values of indicators. For example, a psychiatric diagnosis is based on knowledge of the symptoms. Diagnostic systems such as the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–IV*; American Psychiatric Association, 1994) provide the expert with explicit rules based on the sum score obtained from a list of symptoms, but often the expert does not literally follow such rules but rather relies on tacit indicators. Another example of expert judgment is when people judge themselves on a trait (e.g., "I am shy") or on an attitude (e.g., "I am against capital punishment"). In this case, people are thought to be expert judges regarding more specific, possibly tacit, indicators about themselves that indicate a trait, attitude, or another underlying variable.

One may wonder whether a manifest category resulting from segmentation or expert judgment is in any sense more than an arbitrary segmentation of an underlying dimension. For example, it is a common practice to determine cutoff scores, such as those used to distinguish between depressed and nondepressed persons. The fact that categories are used does not prove that the phenomenon to which the categories refer is category-like. The use of categories may be purely pragmatic. When a category-like manifest variable is used, one must know the nature of this variable to interpret results obtained with it. Manifest categories (e.g., a diagnosis) can correspond to either qualitative differences or quantitative differences. The basic issue is whether the categories at the manifest level are category-like or dimension-like in the latent structure. The complementary issue, whether a manifest dimension (e.g., a sum score) is category-like or dimension-like in the latent structure, is a legitimate question but is not addressed directly here; its answer requires the use of latent class or latent profile models (e.g., see Wilson, 1989, for a discussion). Thus, the present article is asymmetric: Given manifest categories, we attempt to answer whether they are really category-like in the latent structure. If they are category-like at the latent level, then they have the properties of latent categories.

The issue we investigate parallels an issue in cognitive psychology and linguistics, particularly with respect to the nature and meaning of the categories and words we use in daily life. For example, is the concept behind the category *trees* really category-like, or does it correspond better to a dimension of treeness? When the categories are categories of persons—for example, the category of psychiatric patients—and when human cognition intervenes in category assignment (as with expert judgment) then the similarity is even more relevant. The commonalities and differences between our research questions and those of the domain of concepts and categories are illustrative for what we do in this article. In Section 1, we summarize the research on categories and concepts and how it applies to our topic.

Concepts and Categories in Cognitive Psychology and Linguistics

Categories are an important topic of research in cognitive psychology and linguistics. On the basis of empirical evidence, scientists in the domain of cognitive categories believe (a) that cognitive categories cannot be defined in terms of singly necessary and jointly sufficient features, (b) that the distinction between category members and nonmembers is not clear-cut, and (c) that category members differ as to the degree they fit the category, also called typicality (for a summary, see Murphy, 2002). These three conclusions are interrelated and can be summarized in the conjecture that category membership is gradual with no clear cutoff. A similar belief is held in cognitive linguistics (e.g., Lakoff, 1987; Taylor, 1995).

These conclusions contradict the so-called classical (Aristotelian) view. This classical view was described by Rosch (1978) and by Smith and Medin (1981) and defended by Sutcliffe (1993). Wittgenstein (1953) was the first prominent thinker to doubt the classical view. Rosch (1975, 1978) and Smith and Medin (1981) clearly explained and demonstrated empirically why the classical view is invalid. Alternative theories have been developed to explain that people do not use definitions and that categories are gradual instead. These theories are also meant to explain a wide variety of phenomena, such as category decisions, category learning, category-based induction, memory for exemplars, and so on (for overviews, see Komatsu, 1992; Medin & Coley, 1998; Murphy, 2002). We concentrate here on decisions about category membership, because we want to investigate the nature of what we call a *manifest category* based on the attribution of a category label—for example, a personality disorder diagnosis, a self-description as being against capital punishment, or an assignment to a developmental stage.

The first theory states that category membership is derived from the similarity of an element to the prototype of the category. This is the prototype theory. For a description, see Hampton's (1995) study. The similarity is based on a weighted sum of features

present in the element in question. The features and their weights are the content of the prototype. The prototype is of an abstract nature, unless it is instantiated in an extant exemplar. The weighted sum is a continuous variable to be dichotomized for one to decide on category membership or to be used as an input for a choice rule if the decision is between two or more categories.

The second theory states that category membership is determined on the basis of similarity with earlier encountered exemplars from one's (possibly unconscious) memory. This is the exemplar theory. Two well-known elaborations of this view are the context model (Medin & Schaffer, 1978) and the generalized context model (Nosofsky & Palmeri, 1997). It is assumed in these models that for a similarity to be high it needs to be high on all features and that high similarities have a larger weight than do low similarities. The generalized context model is formulated in a rather general way, using various kinds of free parameters, so that it can adapt many phenomena while having similarities to the category exemplars as its core. Empirical comparisons of the prototype theory and the exemplar theory for category decisions tend to favor the exemplar theory (e.g., Medin & Coley, 1998; Murphy, 2002), including when natural categories are studied (Smits, Storms, Rosseel, & De Boeck, 2002; Storms, De Boeck, & Ruts, 2000).

A third theory is not formulated in a formalized way as are the previous two but must be seen as providing an explanation for the shortcomings of these two. This is the knowledge approach theory (Murphy, 2002; Murphy & Medin, 1985) or the explanation-based theory (Komatsu, 1992). In this theory it is stressed that categories are embedded in a broader knowledge about the world and that this knowledge plays an important role in how one deals with and understands categories.

Medin and Coley (1998) and Murphy (2002) noted that an important shortcoming of the prototype theory and the exemplar theory is their neglect of feature relations. That the internal structure of categories has been a neglected topic in the study of categories and concepts is not difficult to explain from the basic conjecture by Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) that categories pick up correlations between features to maximize the informative value of categorization. Categories are clusters of entities based on the correlations between features in a much larger, between-category space. The implication is that categories explain the correlations away (in a statistical sense, not in a causal sense), so that not much correlation is left within the categories. The conjecture of Rosch et al. (1976) is primarily meant for so-called basic-level categories, not for so-called subordinate and superordinate categories. The association of categories with correlated features (in the between-category space) has been empirically corroborated (Devlin, Gonnerman, Andersen, & Seidenberg, 1998; Tyler, Moss, Dunant-Peatfield, & Levy, 2000). Categories defined on the basis of correlated features were found to be more robust against cognitive and neuropsychological deficits—they seem to be stronger categories.

In contrast with the prototype theory and the exemplar theory, an interesting strength of the knowledge approach is that feature relations are recognized—they are part of the knowledge. For example, because one knows that wings help an animal fly, a correlation between wings and flying is a quite natural cognition. This correlation is based primarily on between-category differences. Some categories of animals fly and have wings (various kinds of insects, bats, etc.), whereas other categories of animals do not fly and do not have wings either (elephants, spiders, snails, humans, etc.). But within-category correlation is also no problem for the knowledge approach. For example, for vegetables there is a correlation between being green and growing above the ground. The correlation is not perfect (for example, if one counts tomatoes as vegetables), but the exceptions are rare. Basic biological knowledge can explain the correlation between the green color and growing above the ground. The role that feature relations play in a knowledge approach is that they are quite natural and explained from knowledge one has about the world. No formal theory about feature correlations is developed within the knowledge approach, however, perhaps because there is no compelling evidence for within-category feature correlations to play a role in explaining typicality and category decisions (Murphy, 2002). The evidence in support of a feature-correlation effect is at best rather weak (Malt & Smith, 1984).

One can conclude from this short overview that categories are considered heterogeneous in two senses: Exemplars differ as to how typical they are of the category (typicality differences), and categories can have an internal structure that strongly deviates from a homogeneous uncorrelated structure (structural differences). The internal structure aspect has been somewhat neglected in the prototype theory and in the exemplar theory, but it is stressed in the knowledge approach and in the more linguistic approaches, such as that of Lakoff's (1987) work. Various kinds of internal structures have been described by Storms and De Boeck (1997): one that corresponds to a chainlike structure as described by Lakoff (1987) and another that corresponds to a within-category dimension-like structure (a triangular structure, as in a Guttman scale).

Although the cognitive nature of categories and the linguistic meaning of lexicalized categories is not the topic of our investigation, the results briefly discussed above are nevertheless important because the ingredients are the same as for our topic of interest. In all of our studies, we have elements (persons) that are categorized (the manifest categories) on the basis of features (the indicators). The ingredients are the same, but our research question is different. We are not interested in the cognitive representation or the semantic structure of the categories but in their formal representation in a category-like or dimension-like structure. The two kinds of structure do not necessarily coincide. The issue we want to formulate more precisely for systematic study is whether manifest categories (categories as assigned) can be represented as nothing more than following from cutoffs along a continuum. It is possible that this formal representation is not reflected in the cognitive representation. It has been speculated, for example, that humans tend to think in terms of internal essences (Medin, 1989), which would tend to predispose them toward category-like mental representations of concepts such as mental disorders, whereas the formal representation is an empirical question that may actually be dimension-like.

An interesting link between our research topic and the one from cognitive psychology is that, for both, two types of continua must be distinguished. The first type describes the typicality differences between category members without an internal structure for the category. To understand the first type, assume that all category exemplars are alike in that they show the category features with a probability of, say, .60 and that the features are uncorrelated. This

is actually in line with the well-known latent class model (Goodman, 1972; Green, 1952; McCutcheon, 1987). All category members are equal at the latent level in that they share common feature probabilities. The implications of the assumptions are independence of features and heterogeneity of the exemplars in terms of the features. The features are independent within the category, because they are realized through a mechanism that is independent from feature to feature, following the assumption above. That the features are uncorrelated also means that the category has no internal structure in the sense of within-category correlations between features.

Looking at the realized features, one will notice that the exemplars are heterogeneous, they have quite different feature patterns, because of the stochastic nature of the feature realization. In fact, the probability for two exemplars to share a given feature is only .36. When a category decision is to be made, one can expect that the exemplars with more of the features (as a stochastic result) will be considered category members with more certainty and that their typicality will be considered higher than that of exemplars with an accidentally lower number of features. The equivalent of this is the posterior probability of class membership given the feature realizations. This posterior-probability continuum does not represent anything in the latent level—it merely picks up a characteristic of the realization of the latent structure. We call the resulting kind of continuum a *purely manifest continuum.* It is the illusory effect of a homogeneous process: the same process that leads to independent features and to a lack of within-category structure.

Remarkably, the kind of categories described above is in line both with the classical view and with the common belief that categories are gradual and have no clear cutoff, depending on the level at which one looks. All exemplars are alike at the latent level, which is in conformity with the classical view, and the exemplars show heterogeneity at the manifest level, which is in conformity with the now-common belief that the classical view is wrong. Only the first of the types of heterogeneity mentioned earlier is realized, however (typicality differences). The aspect that is neglected in prototype theory and exemplar theory is neglected here as well (structural differences). Following the first type of heterogeneity, categories are heterogeneous in that not all exemplars are equally good exemplars but not so far as the (latent) internal structure is concerned.

To understand the second kind of continuum, assume that the exemplars differ in the true probabilities of showing the category features. Suppose the probabilities are again high but that they depend on the exemplar (in the range from, say, .60 to .90) and that the feature-realization mechanism is again independent from feature to feature. The exemplars are now heterogeneous at the latent level, because some have higher feature probabilities than do others. The consequences of these assumptions are correlated features and even more heterogeneous feature patterns. The features are all positively correlated because they all tend to occur more in some exemplars (because of their higher probability) and less in other exemplars (because of their lower probability). These correlations stem from differences in probability, notwithstanding the independence of the realization mechanism, which is called *local independence* or *conditional independence* in the statistical literature and is a basic assumption in most statistical models. The resulting categories now have an internal structure, a one-dimensional structure. When one makes the more realistic assumption that not just the exemplars but also the features have an effect on the probability that an exemplar shows the feature, then the stochastic version of the earlier described triangular structure would be obtained. For example, for psychiatric diagnoses this assumption would mean that some symptoms have higher probabilities than do others. Mild symptoms commonly have a higher probability than do severe symptoms. When patients differ in a systematic way, some patients may have the more severe symptoms as well as the milder ones, whereas others may have only the milder symptoms. A one-dimensional internal structure can be a rather good approximation of reality (e.g., for the borderline personality disorder [BPD]). For example, Sanislow et al. (2002) showed that three latent dimensions underlay the borderline symptoms from the *DSM–IV* but also that the intercorrelations of these dimensions are higher than .90 and can reach even .99.

Three sources of differences come into play when looking at the realized feature patterns. First, the exemplars differ randomly because of the stochastic nature of the feature realization. Second, the exemplars differ systematically because of the level of the generating probabilities. The number of category features an exemplar shows reflects both the stochastic nature of the process and a systematic difference at the latent level. Third, the features can also have an effect. The second and third sources determine the probability a feature has for a given exemplar. This probably reflects something about the exemplar (how high its probabilities are overall) and something about the feature (how common it is). It is then possible to separate and estimate the contribution from the three sources: the stochastic source, systematic differences between exemplars, and systematic differences between features.

This idea of separating and estimating the three parts is exactly the idea behind a model from a quite different domain, IRT, as we explain in the *Formalization of Dimcat* section. The newly derived continuum for the exemplars, their overall level of probability, is no longer a surface continuum or an illusory continuum—it is rooted in the underlying latent structure. The number of features is still a manifest continuum, but now it expresses more than a stochastic mechanism. It also reflects systematic underlying differences between the exemplars—the latent contributions of the exemplars to the feature probabilities. The continuum of the systematic underlying differences is not a manifest continuum but a *latent continuum.* It corresponds to the earlier mentioned second type of heterogeneity: structural differences.

This second formal theory of categories, which implies a latent continuum, is no longer in agreement with the classical view, because the exemplars are no longer homogeneous at either the manifest level or the latent level. The theory is in clear agreement, however, with the now-common belief that cognitive categories are gradual and have no clear cutoff. Furthermore, both types of heterogeneity described earlier are now realized. Categories are heterogeneous not only in that not all exemplars are equally good exemplars but also because of the internal structure. This kind of within-category structure can be linked to the notion of fuzzy categories, as discussed by Haslam and Kim (2002) and as tested empirically with taxometric methods by Haslam and Cleland (2002). It should be clear that what we mean by a latent continuum is variation at the latent level and not just at the manifest level. From the way the fuzziness is created by Haslam and Cleland (2002), it can be concluded that this condition is fulfilled.

Thus, one way of framing the issue of whether a category is basically dimension-like is by asking whether categories have a latent continuum or a purely manifest continuum. The results of the studies on categories and concepts cannot answer this question so far as the cognitive representation is concerned, because, as explained, differences in how good exemplars are as exemplars and other effects can stem from either the stochastic nature of a homogeneous latent process or from a genuinely heterogeneous latent process and a similar stochastic component as for the homogeneous process.

In the next section, a frame of reference is described for what it means (a) for the latent structure behind a manifest category to be homogeneous or heterogeneous and (b) for manifest categories to indicate qualitative differences or quantitative differences. Along with this frame of reference comes an approach for modeling data and deciding in what sense their structure is category-like or dimension-like. Later (in the Three Applications section), we describe three empirical applications to illustrate the approach.

A Frame of Reference for Dimension-Like Versus Category-Like Variables

Dimcat

Latent Heterogeneity Versus Homogeneity Within Manifest Categories

Within-category homogeneity means that all persons from the manifest category have the same location on the latent dimension. They are all equal at the latent level. Put another way, the dimension is collapsed to a single point (as far as individual differences are concerned). This latent homogeneity does not prevent heterogeneity at the manifest level of observed indicators, given that the realization of the indicators from the latent location is a stochastic process, in agreement with the assumption that indicators are random variables. *Within-category heterogeneity* means that different persons from the manifest category have different locations on the latent dimension. The distinction corresponds to the distinction between a purely manifest continuum and a latent continuum, as discussed in the previous section. A purely manifest continuum corresponds to homogeneity at the latent level and to a manifest category without internal structure, whereas a latent continuum implies heterogeneity at the latent level and implies that there is internal structure. In fact, three degrees of heterogeneity can be distinguished: manifest homogeneity (as in the classical view on categories), manifest heterogeneity with latent homogeneity (as in the case of categories without internal structure), and manifest heterogeneity with latent heterogeneity (as when the categories have an internal dimension-like structure). Homogeneity at the manifest level can be excluded as unrealistic, so that when we contrast heterogeneous and homogeneous manifest categories, we always refer to their homogeneity and heterogeneity at the latent level. We consider homogeneity more category-like than is heterogeneity and consider heterogeneity more dimension-like than is homogeneity.

Latent Quantitative Versus Qualitative Differences Between Manifest Categories

If manifest categories do not show between-category differences, then there is no reason to distinguish them. Therefore one should assume that they show manifest between-category differences. Regarding the within-category differences, one must differentiate between the case of heterogeneity and the case of homogeneity. (We now use these notions in their latent sense.) When the categories are heterogeneous and a latent dimension suffices to describe the heterogeneity within categories, then we would have *qualitative differences* when the latent dimension differs for members of different manifest categories. As explained above, dimensions are anchored in indicators, and they differ from one another if the discriminations or locations of the indicators are different. Taking personality disorders as an example, suppose that the difference is that the borderline symptoms define a dimension within the borderline category that is different from the dimension that the same borderline symptoms define in the histrionic category. This would mean that for the same symptom, the dimension has another weight depending on the manifest category or that more of the dimension is needed in one category than in another to have the same probability to show the symptom (or to be assigned the symptom). Then one can reasonably claim that the BPD is qualitatively different from the histrionic personality disorder (HPD), because the borderline dimension differs depending on the diagnostic category under consideration. The same would follow if the histrionic symptom dimension differed for persons with HPD and those with BPD. This principle can be generalized to a joint set of symptoms and a two-dimensional structure.

When the manifest categories are homogeneous, the qualitative differences cannot concern the discrimination of the indicators, because there is nothing to discriminate within the category. Only the indicator locations remain as a potential source of qualitative differences. Given that the locations refer to the levels of the indicators, qualitative differences imply that the indicator level profiles differ from one manifest category to another in more than just the overall level. For example, the symptom profile of the histrionic personality disorder may differ from that of the BPD in a qualitative way and not just with respect to its overall lower level of borderline symptoms.

In the case of within-category heterogeneity, *quantitative differences* between manifest categories mean that the latent dimension is the same (same discriminations and/or locations) when applied to members of different manifest categories and that the distribution of one manifest category is located at a lower level than is the distribution of the other category on the same dimension. In the case of homogeneity (no variance in person locations), *quantitative differences* mean that the common category level of the indicator profiles differs depending on the manifest category. For example, it would be reasonable to expect that the preponderance of borderline symptoms is higher in the borderline category than in the histrionic category. The difference is that the quantitative differences can be explained as one manifest category having more or less of the same thing as the other, whereas qualitative differences never can be explained in this way. Qualitative differences concern the anchoring of dimensions with indicators (with respect to discriminations and/or locations): Differently anchored dimensions are different. Considering the contrast between quantitative and qualitative between-category differences, one may consider qualitative differences more category-like than quantitative differences and quantitative differences more dimension-like than qualitative differences.

These two contrasts—heterogeneity versus homogeneity and qualitative versus quantitative differences—can be crossed, as in Figure 1, to make a 2 × 2 classification. This classification is the framework that we use below to explicate the relation between a category-like versus dimension-like latent structure for manifest categories.

In the upper-left panel of Figure 1, two different latent dimensions are shown, one for each of two different heterogeneous manifest categories. The heterogeneity is represented with a normal distribution for each category, although normality of the distributions is not required. In the upper-right panel, the heterogeneity is represented along one common latent dimension. The difference between the two manifest categories is either large (and abrupt) or small (and smooth), as we explain in the next section. In the lower-left panel, two different latent dimensions are again shown, one for each manifest category, but here there are no individual differences within the manifest categories. The within-category homogeneity is represented with a narrow bar. Finally, in the lower-right panel, the two manifest categories are again located along one common latent dimension, but here the two manifest categories are homogeneous, as represented with two bars. Given that in the two lower panels the manifest categories are homogeneous, the between-category differences are abrupt, as we explain in the next section.

Abrupt Versus Smooth Differences

Although we consider the previous two contrasts as the most important, a third contrast, abrupt differences versus smooth differences, can be defined. This contrast cannot be crossed with the other two, as is shown in Figure 1. *Abrupt differences* are qualitative or quantitative discontinuities from one manifest category to another. Differences are necessarily abrupt when the differences are qualitative or when the manifest categories are homogeneous (and therefore do not overlap). Within-category homogeneity as well as between-category qualitative differences imply abrupt differences. *Smooth differences* are necessarily quantitative differences. Only within the combination of within-category heterogeneity and between-category quantitative differences can both smooth and abrupt differences occur between manifest categories. The within-category heterogeneity is indicated by a distribution of persons within the manifest category (e.g., a normal curve, as in Figure 1), and the between-category quantitative differences are indicated by the fact that the two manifest categories can be located at different points along the same dimension. In the example above, if there were no overlap between persons with BPD and those with HPD when they were located on the borderline dimension, then this would be a clear example of an abrupt difference. One would consider this as evidence that persons with BPD are a different latent category from the persons with HPD. A great deal of overlap between persons with BPD and persons with HPD on the borderline dimension, and especially the absence of bimodality, means that the difference between the two manifest categories is smooth. In the case of smooth differences, the two manifest "categories" do not seem very category-like with respect to the dimension under consideration. No overlap and bimodality would indicate abrupt differences. Abruptness implies discontinuity (e.g., Wilson, 1989). Abrupt differences are more category-like and less dimension-like than are smooth differences.

In the upper-right panel of Figure 1, two pairs of normal distributions are shown. The distributions on the left are rather far apart—far enough for the distributions to result in a bimodal distribution when they are added into a joint distribution. The distributions on the right are close enough to result in a unimodal joint distribution.

The contrast between smooth versus abrupt differences cannot be considered a fundamental dichotomy compared with the

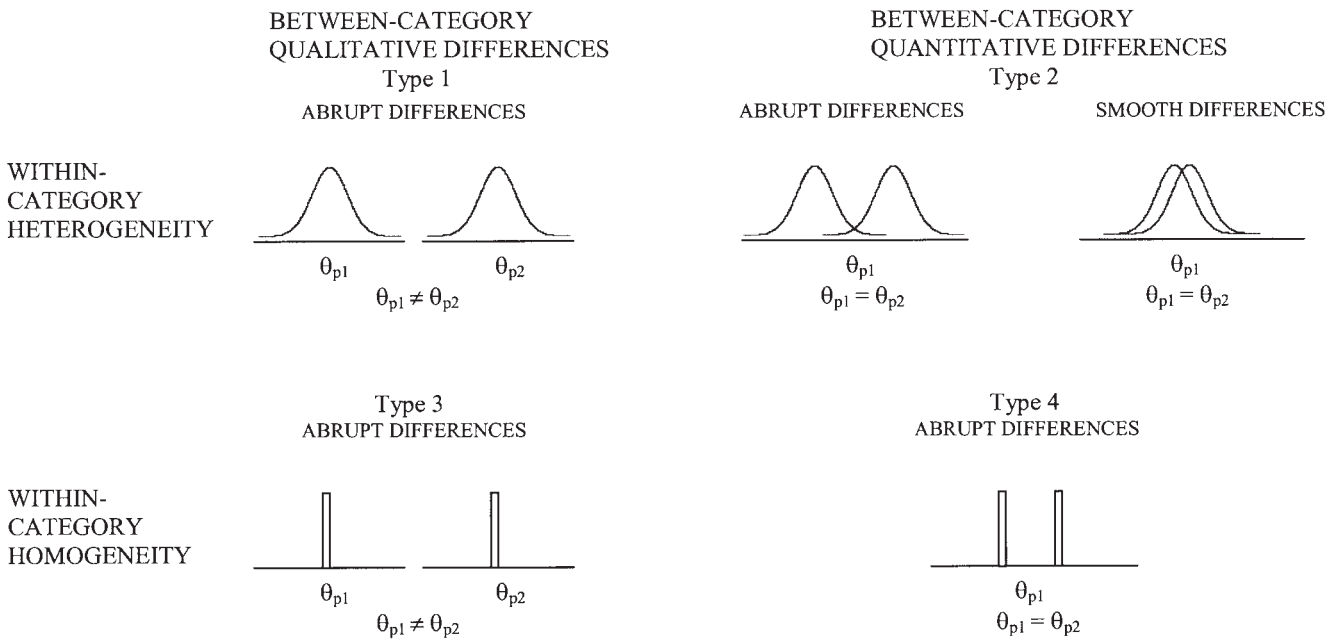

Figure 1. Graphical representation of the 2 x 2 classification of structure.

heterogeneous–homogeneous and qualitative–quantitative dichotomy. The smooth–abrupt dichotomy is relevant only for quantitative between-category differences and is based entirely on the size of the difference between manifest categories and their within-category standard deviations; it is therefore of a gradual kind. But even when there is a gap between the distribution of two manifest categories along the latent continuum, the distinction between the two is purely quantitative and can be expressed in terms of more or less of the same thing.

Simple Versus Complex Qualitative Differences

The qualitative differences between manifest categories can be either simple or complex. Figure 2 shows, for two manifest categories, the range from complex differences to simple differences to no differences. The qualitative differences are complex when each single indicator has a different location, so that the relation looks like an uninterpretable hodgepodge, as in the left panel of Figure 2. Alternatively, simple differences are differences that can be captured with a few parameters. For example, perhaps the location of a few indicators has shifted relative to the location of the other indicators. Suppose that identity disturbance as a borderline symptom is relatively predominant among borderline patients in comparison with other borderline symptoms but that it drops down in the rank order of borderline symptoms when histrionics are considered. Such simple shifts or jumps in the indicator locations are called a *saltus* (Wilson, 1989; see middle panel in Figure 2). Saltus was originally a model for discovering latent classes that explain jumps in indicator locations (betas) from one latent class to another, but a model using manifest categories, the manifest saltus model, has also been developed (Wilson, 1993) and will be used here. For example, suppose there are four indicators, with locations β_1, β_2, β_3, and β_4 on the dimension in the first manifest category and with locations β_1, β_2, $\beta_3 + \delta_{12}$, and $\beta_4 + \delta_{12}$ on the dimension in the second manifest category, with δ_{12} denoting the jump that Indicators 3 and 4 make when going from the first to the second category. In a similar way, shifts can occur in the indicator discriminations. Finally, when there are no qualitative differences, the locations (and discriminations) are the same in the two manifest categories, as in the right panel in Figure 2.

Formalization of Dimcat

For the formal representation of Dimcat, we use symptoms and diagnoses of personality disorders for illustration. The data from which to start are the observations of indicators (e.g., ratings of borderline symptoms) and a manifest category (e.g., the diagnosis of BPD). Most often the indicators are also category-like, and often they are binary, as when symptoms are judged to be present or absent, when a response is correct or incorrect, or when a response is "agree" or "disagree." An extension to polytomous cases is also possible, as discussed below.

The notation for raw scores is as follows:

$$X_{pik} = 0, 1,$$

with

$p = 1, \ldots, P$ (an index for the persons),
$i = 1, \ldots, I$ (an index for the indicators), and
$k = 1, \ldots, K$ (an index for the manifest category to which a person belongs).

When the indicators are symptoms, $X_{pik} = 1$ means that person p from category k is attributed indicator i. The notation for the manifest categories is $C_p = k$, meaning that person p is assigned to manifest category k. In this model, persons are nested within manifest categories.

The manifest category, C, can be a random variable, or it can have fixed values. In a similar way, the parameters from the model to be presented can be either random or fixed. By convention, in formulas we do not condition on C or on parameters, as the conditioning makes sense for random variables only. The fact that the formulas are not given in their conditional format does not imply, however, that C or one or more parameters cannot be random variables.

Building a Generic Formula

All models we describe are models for the probability of a positive (1) response about person p from category k on dichotomous indicator i (based on self-description or other-description): $P(X_{pik} = 1) = 1 - P(X_{pik} = 0)$. The models all share the characteristic that probabilities of this type are a function of indicator parameters such as locations. $P(X_{ijk} = 1) = f(\beta_{jk})$, with β_{jk} being the parameter of indicator j for category k. A common type of function for binary variables is the logistic function, so that

$$P(X_{pik} = 1) = \exp(\beta_{ik})/(1 + \exp(\beta_{ik})) \quad (1)$$

or

$$\eta_{pik} = \beta_{ik}, \quad (2)$$

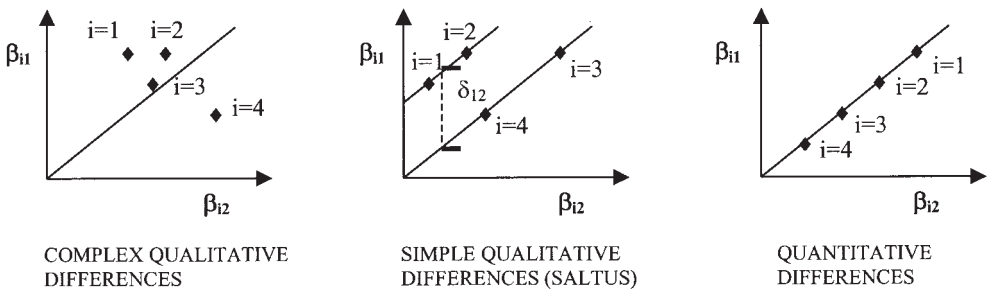

Figure 2. Graphical representation of different kinds of qualitative differences versus quantitative differences.

where $\eta_{pik} = \log(P/Q)$, $P = P(X_{pik} = 1)$, and $P = 1 - Q$.

It follows from Equation 2 that the betas are nothing more than logistic transformations of the probabilities of showing symptom i in category k: $\beta_{ik} = \log(P/Q)$. These logistically transformed probabilities can be represented as the locations of the indicators that can function as anchors on a possible latent dimension. Thus far, all persons have the same set of probabilities. The betas are also called *prevalences,* as they indicate the occurrence of symptoms.

Person differences can be introduced into Equation 1 by substituting $\theta_{pk} - \beta_{ik}$ for β_{ik}, with θ_{pk} denoting the parameter of person p from category k—this locates the persons on the same scale as the indicators; for example, this locates the patients on the same scale as the symptoms, so that the difference between the location of person p and indicator i determines the probability of a response of 1 for a person in category k:

$$P(X_{pik} = 1) = \exp(\theta_{pk} - \beta_{ik})/(1 + \exp(\theta_{pk} - \beta_{ik})) \quad (3)$$

or

$$\eta_{pik} = \theta_{pk} - \beta_{ik}. \quad (4)$$

Because of the inclusion of a theta parameter, the values of the betas are identified only up to an additive constant—one can add a constant to all betas on the condition that the same constant is added to all thetas. Note that the minus sign in Equations 3 and 4 is in a way arbitrary—it could as easily be a plus sign, but the minus sign is the usual convention.

For the example of personality disorders, θ_{pk} reflects the severity of person p on the latent dimension as it applies to diagnosis k, and $-\beta_{ik}$ reflects the prevalence of symptom i for diagnosis k. They both contribute to η_{pi} and to the probability of a 1. One further complication is that the severity is not equally important for all symptoms. To reflect this difference, we adapt the equation as follows:

$$P(X_{pik} = 1) = \exp(\alpha_{ik}\theta_{pk} - \beta_{ik})/(1 + \exp(\alpha_{ik}\theta_{pk} - \beta_{ik})), \quad (5)$$

or

$$\eta_{pik} = \alpha_{ik}\theta_{pk} - \beta_{ik}, \quad (6)$$

where α_{jk} denotes the weight of θ_{pk} in determining the probability of the X_{pik} values.

Equations 5 and 6 represent the two-parameter logistic (2PL) model (Birnbaum, 1968) for a manifest category k; this is the most general model that we use to illustrate Dimcat. Note that in the formulation of the 2PL as in Equation 6, for each indicator a category-specific linear regression equation is obtained with the underlying within-category dimension as a predictor, with α_{ik} as its weight and with β_{ik} as an intercept. See the Appendix for an alternative parameterization.

All other models that we consider follow from restrictions on Equation 6. This model is general in the sense that it can generate all cases in the framework of Figure 1 by choosing the appropriate restrictions.

Descriptive Dimensions

We use the term *descriptive dimension* for the location of the indicators on a dimension. A descriptive dimension is defined by a set of indicator locations. As such it is neutral with respect to the latent-category versus latent-dimension contrast. It is only when we introduce restrictions on within-category heterogeneity and between-category qualitative differences, and on the equality of indicator locations (betas) and indicator discriminations (alphas) depending on the manifest categories (C_ks), that differences in latent structure are obtained.

Before formulating these restrictions on Equation 6 to obtain distinct latent structures, we make use of that equation to characterize two important but distinct features of a dimension: location equivalence and discrimination equivalence. Both types of equivalence are necessary for two dimensions to be identical (i.e., for dimension equivalence). A latent dimension is defined by the location of the indicators and, if individual differences appear, also by the weights of the indicators. A difference between manifest categories in either the location of indicators or their weights, or in both, means that the dimensions differ, unless the difference can be attributed to varying reliability. This is a special case that we do not discuss here, but it is taken into account in the Appendix.

Equivalent dimensions must have equal locations for the indicators. We call the latter *location equivalence.* Because the location parameters are identified only up to an additive constant, location equivalence refers to equality of the location parameters only up to an additive constant, implying that the differences between the indicator locations on the latent dimension are crucial for location equivalence. If marks on the scale do not correspond, then the meaning of the dimensions also differs. This first aspect of a latent dimension is independent of individual differences among persons.

Equivalent dimensions must have indicators with weights (discriminations) that do not depend on the dimension. Equality of discrimination parameters is called *discrimination equivalence.* If the differentiation capacity of an indicator depends on the manifest category, then the meaning of the dimensions differs between the manifest categories. Note that the discriminations are identified only up to a multiplicative constant: Multiplying the discriminations with a constant is compensated by dividing the variance of the underlying dimension by the squared value of the same constant. This second aspect of a dimension makes sense only if there are individual differences among persons, because the alphas are the weights of latent individual differences (in terms of theta).

The notions of location equivalence and discrimination equivalence are related to the notions of factorial equivalence, measurement invariance, and differential item functioning (DIF). Factorial equivalence is of relevance here, because Takane and de Leeuw (1987) showed that the factor model results when the normal-ogive function is used in place of the logistic function used above and because the two functions are practically identical except for a different slope. Often in factor analysis one is not interested in the means, and the model is then formulated for within-category deviation values (with a mean of zero), so that factorial equivalence is limited to the factor loadings. We refer to this notion of factorial equivalence as *factorial equivalence in the limited sense.* Both Reise, Widaman, and Pugh (1993) and Meredith (1993), however, pointed out that the full factor model includes an explanation for the means, so that factorial equivalence in this broader (and full) sense includes location equivalence as well. Reise et al. (1993) distinguished between full invariance and partial invariance (see Byrne, Shavelson, & Muthén, 1989). Both are related to the

factor loadings, independently of the factor variances and covariances. Full invariance means category-invariant loadings for all variables, whereas partial invariance implies that a substantial amount of the loadings are invariant so that a common metric can still be used. For binary indicators, the factor analytic or structural equation model for binary items would be equivalent to an IRT model but of the normal-ogive type instead of the logistic type (Bock, Gibbons, & Muraki, 1988; Muthén, 1984). A logistic variant is described by McKinley and Reckase (1983).

As to *measurement invariance,* Reise et al. (1993) referred to the same two aspects we have discerned in dimension equivalence. Meredith (1993) started from a definition stating that the cumulative distribution function of the measurement indicators may not depend on external factors beyond the underlying latent variables one assumes to explain the indicators. Simply stated, the measurement of intelligence, for example, may depend only on intelligence and not also on external factors, such as one's ethnicity. Invariance refers to all aspects of the cumulative distribution (expected value, variance, and higher moments) and implies both location and discrimination equivalence.

Lack of location equivalence is called *uniform DIF* in test theory, and lack of both location equivalence and discrimination equivalence is called *nonuniform DIF* (see, e.g., Holland & Wainer, 1993). Methods to detect DIF are described in the literature (Holland & Wainer, 1993; Millsap & Everson, 1993), but some of the DIF tests do not distinguish between unequal locations and discriminations.

In summary, discrimination equivalence refers to the indicator-specific slope of the equation (α_{jk}). Location equivalence refers to the indicator-specific intercept of the equation ($-\beta_{ik}$). Location equivalence is sometimes not investigated in empirical studies in the literature, because the factor model is used not in its full formulation but rather for deviation transformed variables (Reise et al., 1993).

Types of Latent Structure

We present the unrestricted latent structure followed by three constrained latent structures. Note that the structure that is presented here as the unrestricted structure (i.e., the generic) is still restricted, in that it corresponds to a 2PL model for each manifest category. For example, it is assumed that the structure is unidimensional within each manifest category (but not necessarily between manifest categories). Extensions to less-constrained, unrestricted cases (e.g., multidimensionality within manifest categories) are discussed below.

1. In the first type of latent structure (corresponding to the upper-left panel in Figure 1), the latent dimensions are qualitatively different depending on the manifest category, and the persons are heterogeneous within manifest categories. An example is categories of athletes defined on the basis of the kind of sport, with performance levels as indicators. These categories would be between-sports categories with performance indicators. Within each category there are clear and systematic quantitative differences in athletes' performances, and from one category to the other there are qualitative differences in the kind of performances at which the athletes are good.

As far as the modeling is concerned, no restrictions on Equation 6 are introduced, and it is therefore reflected in the general Equation 6, where for k and k' the β_{ik} are allowed to differ from the $\beta_{ik'}$, and the α_{ik} are allowed to differ from the $\alpha_{ik'}$. This first type will serve as the reference type in the presentation of the other types, given that all others can be defined as restrictions on this one. In this first type, there is continuity within each qualitatively distinct category. Because the latent dimension differs depending on the manifest category k, the differences between manifest categories are qualitative. Both the indicator locations (β_{ik}) and the indicator discriminations (α_{ik}) are allowed to be category-specific. Because individual differences among persons, as expressed in θ_{pk}, are allowed, the manifest categories are heterogeneous. A special case is one with category-specific locations but common discriminations. Note that these type of differences would not be identified when factorial equivalence in the limited sense was the only criterion used to detect qualitative differences. In this case, the locations of the indicators are category dependent, but their discriminative power is not category dependent.

2. In the second type of latent structure (corresponding to the upper-right panel in Figure 1), the latent dimensions are quantitatively different depending on the manifest category, and the persons are heterogeneous within manifest categories. An example is the categorization into a professional and a nonprofessional category of athletes within the same sport. One can expect that both professionals and nonprofessionals differ in how well they perform at various contests, but the professionals would be clearly better overall than the nonprofessionals. These manifest categories would be within-sport categories with performance indicators.

The second type of latent structure differs from the first in only one respect: For any pair of manifest categories ($k \neq k'$), the location of the manifest categories may differ only along a common underlying dimension. As a result of the absence of qualitative differences, all betas and all alphas of each of the indicators are equal over manifest categories: $\beta_{i1} = \ldots = \beta_{ik} = \ldots = \beta_{iK} = \beta_i$, and $\alpha_{i1} = \ldots = \alpha_{ik} = \ldots = \alpha_{iK} = \alpha_i$. The second type can be formulated as follows:

$$\eta_{pik} = \alpha_i \theta_{pk} - \beta_i \tag{7}$$

with β_i denoting the common location parameters, with α_i denoting the common discrimination parameters, and with $\mu_{\theta k} \neq \mu_{\theta k'}$, for $k \neq k'$.

Depending on how the manifest categories are distributed along the dimension, the differences between the manifest categories may be abrupt or smooth. There is no clear-cut criterion to distinguish between smoothness and abruptness, but two criteria that are often associated with abrupt differences are lack of overlap and bimodality. As discussed above, these criteria are less straightforward than one might think.

First, much depends on the kind of distribution one wants to assume for the two manifest categories. For example, lack of overlap can also look perfectly smooth, as when persons within each manifest category are distributed uniformly and the two distributions touch but do not overlap. Second, much depends on whether one looks at the manifest level or the latent level. For example, Grayson (1987) showed that depending on the discriminations and on the locations of the indicators, a bimodal distribution of sum scores may result from a unimodal distribution of person locations (thetas). Although Grayson did not demonstrate the opposite—a unimodal distribution of sum scores can result from a bimodal distribution of person locations (thetas)—this is

possible as well. The actual outcome depends on the locations and discriminations.

3. In the third type of latent structure (corresponding to the lower-left panel in Figure 1), the latent dimensions are qualitatively different depending on the manifest category, and the persons are homogeneous within manifest categories. This type of latent structure differs from the first in only one respect: The manifest categories are homogeneous in their latent structure. An example is the categories of athletes defined on the basis of their knowledge of the basic rules of the sport they practice. Within each category there is homogeneous knowledge of the basic rules (they all know the basic rules), although when questioned one may give a wrong answer now and then. The differences between the categories are qualitative in that the athletes differ in the kind of rules they know depending on the sport they practice. These categories are between-sport categories with rule-knowledge indicators.

As a result of the homogeneity restriction, all thetas within the same category are equal: $\theta_{pk} = \theta_{p'k} = \theta_k$ for all pairs of persons p and p' and for all values of k. In this type, the manifest categories do not have any dimension-like character: They are qualitatively different between categories and perfectly homogeneous. There is still an ordering possible for the indicators, but this means nothing more than that the probability of a certain response for a given indicator is different than the probability for other indicators. Note that when there are no individual differences within a manifest category, there is no longer any basis for using a discrimination parameter. The third type can therefore be formulated as follows:

$$\eta_{pik} = \theta_k - \beta_{ik} \tag{8}$$

where for any pair of manifest categories, $k \neq k'$, the β_{ik} may differ from the $\beta_{ik'}$, with θ_k denoting the location of all persons p with $C_p = k$.

4. In the fourth type of latent structure (corresponding to the lower-right panel in Figure 1), the latent dimensions are quantitatively different depending on the manifest category, and the persons are homogeneous within manifest categories. This type of latent structure differs from the first in two respects: The manifest categories are homogeneous (like the third type), and the differences between the manifest categories are quantitative (like the second type). An example would be the categories of persons who do versus do not play chess. Those who play chess would know all the basic rules, and those who do not would also be rather homogeneous in their lack of knowledge. They may guess and be correct on some of the rules, but no major systematic differences would exist. So the difference between the two categories is quantitative. The former category simply has a much higher knowledge than does the latter. These categories are within-sport categories with rule-knowledge indicators.

As a result, all person locations (thetas) within the same manifest category are equal: $\theta_{pk} = \theta_{p'k} = \theta_k$ for all pairs of persons p and p' and for all values of k, as in the third type; all indicator locations (betas) are also equal: $\beta_{i1} = \ldots = \beta_{ik} = \ldots = \beta_{iK} = \beta_i$. Again there is no basis for using a discrimination parameter. In this fourth type, homogeneous manifest categories are located within a latent dimension. The fourth type can be formulated as follows:

$$\eta_{pik} = \theta_k - \beta_i \tag{9}$$

where β_i denotes the common location parameters.

Degrees of Being Dimension-Like Versus Category-Like

Considering being category-like a matter of degree and believing hybrid structures to be common, Waller and Meehl (1998) stated, "Taxonicity does not preclude dimensionality . . . . the convenient dichotomy taxonic-vs.-dimensional should, strictly speaking, read 'taxonic-dimensional vs. dimensional only'" (p. 9). Haslam and Kim (2002) also drew attention to the fact that "matters of kind and matters of degree, itself [might] be a matter of degree" (p. 311), pointing also to an early acknowledgement of this view by Meehl (1979). There are two reasons for thinking of degrees of being dimension-like versus category-like. The first reason is that the features that define the four types of latent structure are crossed, such that some latent structures are defined by some features that are category-like and some that are dimension-like. The second reason is that the features that define the four types of latent structure are often realized only imperfectly and are thus matters of degree.

The only type of latent structure that is thoroughly category-like is the homogeneous qualitative difference structure (Type 3). All other structures are at least partly dimension-like. The heterogeneous quantitative difference structures (Type 2) are thoroughly dimension-like if the differences between manifest categories are smooth. If the differences between manifest categories are abrupt, meaning that each manifest category has a distribution that is different enough along the single latent dimension, then heterogeneous quantitative differences are a hybrid structure. The second type of hybrid structure is the heterogeneous qualitative difference structure (Type 1), which is category-like because the differences between manifest categories are qualitative yet is also dimension-like because there is heterogeneity of persons within a descriptive dimension for each manifest category. The third type of hybrid structure is the homogeneous quantitative difference structure (Type 4), which is category-like because there is homogeneity of persons within the manifest categories yet is also dimension-like because the manifest categories differ as to their locations on a common descriptive dimension.

The features that define what it means to be category-like versus dimension-like can be realized to a stronger or weaker degree. That is, within-category homogeneity, between-category qualitative differences, and abrupt between-category differences can be small or large. First, for within-category homogeneity to be small or large means that the within-category variance is small or large, respectively. Second, what it means for between-category differences to be small or large is simple when the differences are quantitative: Small versus large differences correspond, respectively, to small versus large Cohen's d values (a standardized effect size measure), given that the distributions are normal (or symmetrical). The extent of qualitative differences is more complex. Qualitative differences between manifest categories are complex when they are not restricted to a few indicators or to a few principles. For locations, qualitative differences can be thought of as jumps of indicators on the descriptive dimension when going from one manifest category to another. When the jumps can be summarized with a few saltus parameters (when only a few indicators jump or when groups of indicators each jump over the same distance), the differences are simple. When many saltus parameters are required, the differences are complex. Third, whether abrupt differences are small or large depends on the size of the two previous types of heterogeneity and

on the distributional properties (e.g., bimodality, degree of overlap). We stress here that the differences between the four types of latent structures are gradual and not absolute. This is completely in line with the overall idea behind the framework that being categorical is not itself categorical.

Empirical Methodology, Modeling, and Software

On the basis of Dimcat, one can use empirical procedures to test the category-like versus dimension-like nature of a concept. The observables are indicators and manifest categories. The simplest case is that only one manifest category is considered. This means that only one feature of the framework is relevant: the within-category homogeneity versus heterogeneity. The more complex case is that more than one manifest category is considered. This allows also for investigating qualitative versus quantitative differences and smooth versus abrupt differences as features of the latent structure. For example, persons with BPD could be contrasted with persons with HPD (BPD, HPD), with control participants (BPD, controls), or with both (BPD, HPD, controls).

For the within-category aspects as well as for the between-category aspects, the methodology is necessarily relative, even when the "truth" would be absolute, because the methodology is always limited. The results depend on the choice of indicators and of alternative manifest categories. To study within-category homogeneity versus heterogeneity, one needs a set of indicators. For personality disorder categories, symptoms are an evident choice, but a difficult issue is how one can make sure that all relevant symptoms are included. For other types of manifest categories, it is often less evident of what kind the indicators should be. The choice of indicators (features) is also a difficult issue in the cognitive study of concepts and categories (Murphy, 2002, pp. 45–46). In general, one can never be certain whether the crucial indicators are included in the study. On the other hand, an a fortiori type of reasoning applies. If, for the indicators that are chosen, within-category heterogeneity is found, then one can conclude against homogeneity. The a fortiori argument is that the manifest category will remain heterogeneous when other indicators are added. If, however, a manifest category turns out to be homogeneous, then the conclusion can change if other indicators are added, given that these new indicators may reveal the heterogeneous nature of the manifest category.

For the between-category aspects, the conclusions may also depend on the choices one makes. The latent structure may be category-like in contrast with one alternative manifest category (for example, persons with BPD contrast with control participants) but not in contrast with a second alternative manifest category (for example, persons with BPD in contrast with persons with HPD). When there is more than one manifest category, another complication is which indicators one should consider: indicators of one of the manifest categories (and of which one?) or indicators of all manifest categories. For example, one may investigate the manifest categories BPD versus HPD with a set of borderline symptoms as indicators, with a set of histrionic symptoms as indicators, or with a set that comprises both. The result may depend on which set of indicators is being used.

One cannot give an absolute answer to the general question of whether the BPD is category-like or dimension-like, because the answer may depend on the methodology: the indicators and the other groups one wants to consider (e.g., control participants? which other personality disorders?). Thus, being category-like is not only a matter of degree, it is also relational. If one manifest category is considered in isolation, then the relational character is less evident, because as soon as heterogeneity is found, the conclusion is that the manifest category is heterogeneous. If, however, a manifest category is studied in the context of other manifest categories, then the position on the horizontal axis of the framework depends on which other manifest categories are considered. It may turn out that a Diagnosis A shows qualitative differences with Diagnosis B but only quantitative differences with Diagnosis C. A general conclusion is not possible in that case; only a relative one is possible: In relation to Diagnosis B, Diagnosis A shows qualitative differences, but this is not true in relation to Diagnosis C.

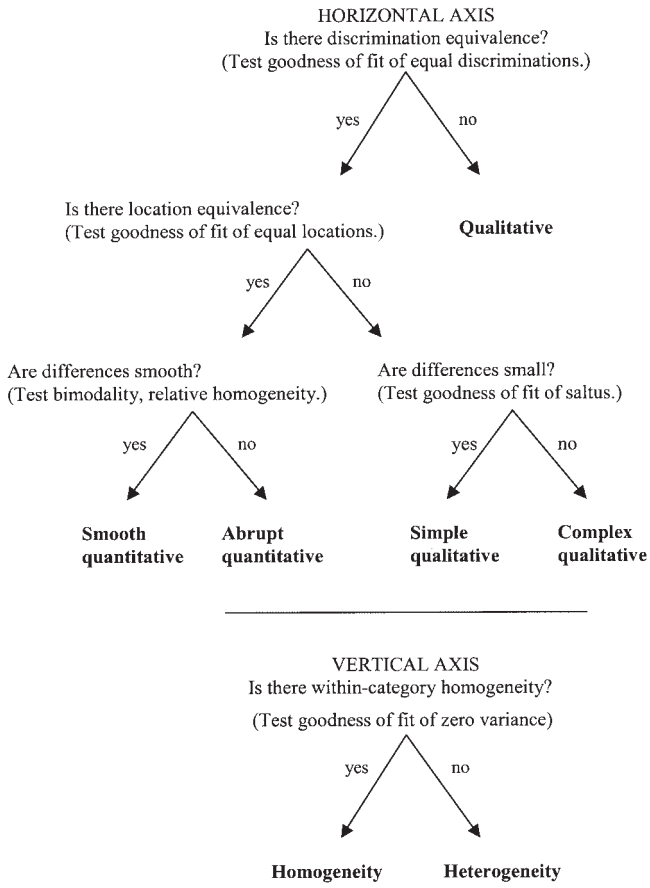

Figure 3. Flow chart for assessing Dimcat distinctions.

Modeling

To find out which type of latent structure applies, one should distinguish the horizontal and vertical axes of Figure 1. The structure is less category-like and more dimension-like when going up on the vertical axis and when going to the right on the horizontal axis. A flow chart for assessing distinctions in the framework appears in Figure 3.

The horizontal axis: Quantitative differences versus qualitative differences. Qualitative differences can be of two types: differences in discrimination and differences in location. Discrimination equivalence and location equivalence are two ways in which qualitative differences can be lacking and are restrictions on Equation 6 (or Equation A1 in the Appendix).

The following order of analyses is proposed. First, we estimate without any restrictions the general model of Equation 6 as parameterized in Equation A1 of the Appendix. This model is called QUAL1&2–HET because it describes qualitative differences of both types and because the manifest categories are allowed to be heterogeneous. Second, the discriminations are restricted to be equal over the manifest categories (discrimination equivalence), yielding model QUAL2–HET, because qualitative differences are allowed only for the locations. The QUAL1&2–HET and QUAL2–HET models are variants of a Type 1 structure. Third, the locations are restricted to be equal over manifest categories (location equivalence), yielding model QUAN–HET, because only quantitative differences remain. The QUAN–HET model is a Type 2 model.

Heterogeneous quantitative differences (Type 2) are nested within heterogeneous qualitative differences (Type 1); and, within Type 1, QUAL2–HET is nested within QUAL1&2–HET. We have chosen to estimate three models of decreasing complexity, QUAL1&2–HET, QUAL2–HET, and QUAN–HET, omitting the fourth possible model, QUAL1–HET, a model with location equivalence without discrimination equivalence. We believe it makes sense to restrict the discriminations first, because their estimation is less reliable than the estimation of the locations. Models QUAL1–HET (which we did not test) and QUAL2–HET are not nested in one another.

The vertical axis: Heterogeneity versus homogeneity. As for the investigation of heterogeneity versus homogeneity, with the restriction of θ_{pk} to have zero variance for all values of k, models that parallel the heterogeneous ones are obtained: QUAL1&2–HOM, QUAL2–HOM, and QUAN–HOM, but QUAL2–HOM cannot be distinguished from QUAL1&2–HOM. The homogeneous models are nested within their heterogeneous counterparts. Homogeneous qualitative differences (Type 3) are nested within heterogeneous qualitative differences (Type 1), and homogeneous quantitative differences (Type 4) are nested within heterogeneous quantitative differences (Type 2). Note that it is possible that one of the manifest categories is homogeneous and the other is not. This is not a serious complication, as it would mean for example that $\sigma_1^2 \neq 0$, whereas $\sigma_2^2 = 0$, which is a less severe restriction than when both variances are restricted to zero.

In a preliminary and exploratory investigation of heterogeneity, one can use an internal-consistency index, Cronbach's alpha, in each manifest category. High values of this coefficient are an indication of heterogeneity. Low values, however, can have two, possibly combined, causes: low heterogeneity and multidimensionality. Cronbach's alpha can be tested for statistical significance and thus can also be used in a hypothesis-testing approach.

Smooth versus abrupt differences. To distinguish within the top-right panel of Figure 1 between smooth versus abrupt differences, we plot the distributions of the theta in the different manifest categories to inspect the joint distribution for multimodality. One should not look only at the plots without also taking the model estimation into account, however, because what appears as smooth may actually be a discrete latent process, as we illustrate in a simulation study reported in *Application 2: Attitudes Toward Capital Punishment.*

Simple versus complex qualitative differences. To distinguish within the top-left panel of Figure 1 between simple versus complex qualitative differences, we investigate whether the lack of location equivalence can be reduced to a few saltus parameters. In principle this method could be followed for discriminations as well as for locations, although it was originally formulated for locations, but it turned out that in our applications, discrimination equivalence was a tenable assumption.

Statistical approaches to testing. A first aspect of testing is whether a model fits the data in an absolute sense, independently of a comparison with other models. We follow two approaches to deal with this problem. In one application, a bootstrap method is used, and in the other applications, a Pearson χ^2 test is used for an equivalent conditional maximum-likelihood (CML) formulation of the selected model, because the CML framework has nicer statistical properties when it comes to testing absolute goodness of fit (Glas, 1988). Given that the issue here is to select the best-fitting model to identify the most appropriate latent structure (Type 1, 2, 3, or 4), the absolute goodness of fit is less important than is the relative goodness of fit.

Second, a broad range of methods is available to test relative goodness of fit. The first kind of test is the likelihood ratio test. This test is based on –2logL (L for likelihood), also called the *deviance.* The test compares the deviance value of two models, one of which is nested into the other. The difference of the two deviances is chi-square distributed with a number of degrees of freedom equal to the reduction in the number of parameters of the nested model. Unfortunately, this test is no longer valid if one or more of the restrictions includes a boundary value, such as a variance that is fixed to zero. If the test is used nevertheless to test such zero-variance models, then the result is conservative (Verbeke & Molenberghs, 2000). We used the conservative test, as it did not make a difference in the applications whether the correct or the conservative test was used. We used the regular likelihood ratio test for the horizontal axis, but for the vertical axis (to distinguish between heterogeneity and homogeneity), we used the conservative test. We also used the regular likelihood-ratio test for simple versus complex qualitative differences, because the saltus models are a reduced form of general qualitative differences.

An important problem with model selection is that the more complex models by definition have a higher chance to fit the data, whereas the simpler models are more parsimonious. A good balance of the two qualities is desirable. This explains the popularity of so-called information criteria. The Akaike information criterion (AIC; Akaike, 1973) or Bayesian information criterion (BIC; Schwarz, 1978) can be used to compare models while taking their complexity into account. Both the AIC and BIC penalize models for a higher number of parameters. The penalization is more severe in the BIC, because it increases with the log of the number of persons; therefore the BIC tends to favor the simpler models more than the AIC does, especially for a large sample size. For both the AIC and BIC, lower values indicate better model fit.

It is also possible to test individual parameter values against their null hypothesis value using Wald tests—dividing a parameter estimate by its standard error. The resulting statistic follows a t distribution, but for a high number of observations (as in our applications) it can be interpreted as a z distribution (as asymptotically it is). Like the likelihood ratio test, Wald tests are conservative for the null hypothesis of zero variance.

As to the differentiation of the various types of structure, one should realize that this is an issue not specific to our approach, given that we use only extant item response models. A complicating factor is that the difference between the various structures is gradual, as we have explained, so that by definition the differentiation power will be small when the differences are small. Nevertheless, we have conducted a simulation study with 40–80 data sets per type of differentiation (Hidegkuti & De Boeck, 2004), and it was found that for the likelihood ratio test, AIC, and BIC, the differentiation power was very good for all but two differentiations, even when small data sets were used (2×100 respondents and just 10 indicators). The two more problematic cases were the following: (a) The one-parameter logistic model was preferred in about 35% of the cases when the 2PL model was the true model, and (b) discrimination equivalence was preferred over lack of discrimination equivalence in 30% of the cases in which the true model violated the equivalence. The two differentiations in the other direction did not yield any problems. When the standard deviation of the degree of discrimination was raised from .10 to .25 (so that the one-parameter logistic model was violated to a larger extent and the lack of discrimination equivalence was stronger), however, these two differentiations were no longer problematic. We also performed a specific simulation study for Application 2, because the difference between within-category homogeneity and within-category heterogeneity cannot always be distinguished by visual inspection of the histogram. We report the results with Application 2, and they confirm the differentiation power of our modeling approach. Finally, Mislevy and Wilson (1996) also reported simulation results regarding the saltus model.

Software

Two kinds of software are available for testing distinctions in Dimcat: general statistical software and IRT-specific software. To estimate indicator parameters while getting rid of person parameters, some programs assume a particular distribution of persons, usually a normal distribution (i.e., they use marginal maximum-likelihood estimation; e.g., Adams, Wilson, & Wu, 1997; Mislevy, 1984), whereas other programs make no assumptions about the distribution of persons (i.e., they use CML estimation). Midway between is a histrogram distribution, which is very flexible (Adams et al., 1997).

The software available for model estimation with marginal maximum-likelihood includes general statistical software for nonlinear mixed models—for example, SAS PROC NLMIXED (SAS Institute, 1999)—and IRT-specific software such as BILOG (Mislevy & Bock, 1989), MULTILOG (Thissen, 1997), and CONQUEST (Wu, Adams, & Wilson, 1998). An alternative is loglinear modeling, which uses CML estimation of indicator parameters. Using CML, the general program LOGIMO (Kelderman & Steen, 1993) can perform IRT loglinear analyses. The IRT-specific program OPLM (Verhelst, Glas, & Verstralen, 1994) is also based on CML. Both programs allow for a priori differences in indicator discriminations but not for the estimation of discrimination parameters. In the absence of a theory that specifies discrimination values a priori, such methods as preexploring the data (OPLM includes a subroutine for this purpose) could result in good approximate discrimination values.

Given that SAS is a widely used software package, we used SAS PROC NLMIXED (SAS Institute, 1999). This procedure was developed for nonlinear mixed models (McCulloch & Searle, 2001). The IRT models we described are of this type (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). Our models are nonlinear in two ways: because of a nonlinear link function (e.g., a logistic function or a normal-ogive function) and because they are not linear in the parameters, as when products of parameters appear in the model (as in $\alpha_{ik}\theta_{pk}$). The models are mixed because they contain fixed effect parameters as well as random-effect parameters. The alphas, betas, and gammas (see Appendix) are fixed-effect parameters in that they do not vary at random over individuals, but θ_{pk} is a random-effect parameter. The nonlinear mixed models are generalizations of linear regression models. SAS provides not just the logistic variants of the models but also the normal-ogive variants, so that the factor-analytic versions of the models can also be estimated. We show in the Appendix how the estimation of models based on Equation A1 can be set up in SAS PROC NLMIXED. For more information, one can also consult Appendixes A and B in Rijmen et al.'s (2003) article.

Extensions

The Dimcat framework can be extended in at least three ways.

The first extension is to allow for multidimensionality within manifest categories. This requires that θ_{pk} be given a dimension index: θ_{pkr} , $r = 1, \ldots, R$. Note that as presented the framework already allows for multidimensionality between manifest categories (such a structure would fall on the left side of Figure 1). To deal with multidimensionality within manifest categories, one either assigns indicators to specific dimensions, or one estimates the discriminations of indicators on each dimension (using dimension-specific weights, α_{ikr}, with r indicating the dimension: $r = 1, \ldots, R$). In the latter case, the problem of unreliable estimates of discriminations becomes serious, because there are now K sets of discriminations per manifest category, and possibly $K \times R$ sets for the total. We are not interested in the exact values of the alphas, however, but in the test of whether the equality constraint on the alphas makes a difference.

The second extension is to allow for polytomous indicators (instead of only binary indicators). Although several models for polytomous variables can be incorporated into the framework, robustness of estimation is improved when the structure of the indicator response categories is constrained. For example, in the rating scale model (Andrich, 1978), the steps from one category to another do not depend on the indicator, but in the partial-credit model (Masters, 1982), a different location is specified for each response option within each indicator.

The third extension is to allow for latent categories (instead of only manifest categories). Latent categories cannot be identified simply on the basis of manifest variables. This extension implies a reformulation of the models in terms of latent classes (Mislevy & Wilson, 1996; Rost, 1990, 1991; Wilson, 1989). The latent classes

do not necessarily correspond to the manifest categories—that is, the latent classes approach does not guarantee that the categorical variable of interest will emerge. Consequently, issues regarding the manifest categories cannot be dealt with directly. Furthermore, because latent classes are not defined a priori, they require interpretation before they can be labeled. A generalized approach to formulating such problems was described by Pirolli and Wilson (1998). As we discuss later in this article, the well-known taxometric approach is directed to latent categories while concentrated mainly on one feature of our framework.

Except for the latent class extension, the extended models can in principle be estimated with SAS PROC NLMIXED, but in practice a model with a high dimensionality will prove difficult to estimate. Other IRT software is also available, but it would lead us too far afield to give an overview, and high dimensionality is also a problem for those programs. We do not dwell on Bayesian methods (e.g., Beguin & Glas, 2001; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000), because they are not broadly accessible to researchers in psychology and because for high dimensionalities they require a large sample size.

Classical Methods to Distinguish Between Qualitative Differences and Quantitative Differences

Instead of using an IRT approach, as we presented, one can concentrate on other methods to distinguish between category-like and dimension-like latent structures. An early and popular method for distinguishing qualitative differences from quantitative differences was checking for multimodality at the manifest level. If two or more manifest categories are investigated and the joint distribution of the sum scores has multiple modes corresponding to the different manifest categories, then this is considered a clear sign that the manifest categories are qualitatively different. This method has often been applied to investigate the category-like nature of personality disorders (e.g., Kass, Skodol, Charles, Spitzer, & Williams, 1985; Livesley, Jackson, & Schroeder, 1992; Nestadt et al., 1991; Zimmerman & Coryell, 1990). In none of these studies was any evidence found for multimodality. As explained above, this criterion is equivocal. Multimodality shows only that there are large between-category differences at the manifest level, but the difference at the latent level can be either quantitative or qualitative—and, if quantitative, multimodality does not necessarily apply to the latent level. Alternatively, lack of multimodality can occur when the differences between manifest categories are qualitative. One reason for the popularity of multimodality may be the implicit assumption that multimodality at the manifest level was induced by multimodality at the latent level. As discussed above (cf. Grayson, 1987), this assumption may be mistaken.

A second method for distinguishing qualitative differences from quantitative differences is checking factorial equivalence in its limited sense across manifest categories. If, in different manifest categories, the same factor loadings are found, then it is concluded that the latent structure is dimension-like. This method has been applied quite often in the study of personality disorders, with the result that a dimension-like structure seems appropriate (e.g., Livesley & Schroeder, 1990; Livesley, Schroeder, Jackson, & Jang, 1994; Tyrer & Alexander, 1979). From the approach we have developed, however, it is clear that factorial equivalence in its limited sense is important but also that it is only half of the story. Strict factorial equivalence as defined by Meredith (1993) is required.

A third method for distinguishing qualitative differences from quantitative differences is the taxometric approach developed by Meehl (1973, 1995, 1999, 2004). Although the underlying model is not based on manifest categories, data from persons belonging to different manifest categories are often used in its application. Taxometric methods have been applied to many psychological variables, including BPD (e.g., Rothschild, Cleland, Haslam, & Zimmerman, 2003; Trull, Widiger, & Guthrie, 1990), dissociation (e.g., Waller, Putnam, & Carlson, 1996; Waller & Ross, 1997), worry (e.g., A. M. Ruscio, Borkovec, & Ruscio, 2001), depression (e.g., Haslam & Beck, 1994; A. M. Ruscio & Ruscio, 2002; J. Ruscio & Ruscio, 2000), sexual orientation (e.g., Gangestad, Bailey, & Martin, 2000; Haslam, 1997), and personality (e.g., Gangestad & Snyder, 1985, 1991; Strube, 1989). The main findings were summarized by Haslam and Kim (2002), who concluded that several psychopathological variables are *taxonic* (the term used in taxometrics for category-like), such as schizotypy and antisocial personality disorder (APD), whereas other variables are *nontaxonic* (dimension-like), such as depression. As for personality variables, Type A personality seems taxonic, whereas the five-factor model traits and the Jungian traits seem nontaxonic.

The taxometric method MAXCOV (Waller & Meehl, 1998) is based on two assumptions: (a) Between latent categories the indicators are correlated, and (b) within latent categories the indicators are not correlated. Suppose that there are two latent categories represented in a sample and that they have an overall effect on the indicators. Then, as a consequence of the two assumptions, the sum of the indicators can be a good indicator of category membership. Persons with high sum scores will belong mostly to one category, and persons with low sum scores will belong mostly to the other category. On the other hand, persons with moderate sum scores can come from both categories. Therefore, it is expected that the covariance between pairs of indicators will show a curvilinear relation with the sum score of the remaining indicators. In practice, the sum score is divided into intervals, and the covariances are determined for pairs of indicators within each interval. The interval with the maximum covariance (MAXCOV) is the HITMAX interval. If the curve is flat, then the conclusion is that the latent structure is not category-like but dimension-like. Note that in correspondence with the distinction that was made earlier, the manifest categories do not play any role in the method, except to determine the samples.

Taxometric methods were later extended from a pairwise approach to a multivariate approach. Either the first eigenvalue in a principal-components analysis is used as a criterion instead of the covariance between pairs of indicators (the MAXEIG method; Waller & Meehl, 1998) or the distribution of factor scores on the first factor is checked for multimodality (the L-Mode method; Waller & Meehl, 1998). Waller and Meehl (1998) showed that the MAXCOV, MAXEIG, and L-Mode methods are formally equivalent for the case of homogeneous taxa.

The taxometric approach focuses on whether taxa are homogeneous, which corresponds to the lower portion of Figure 1 (i.e., the types that show within-category homogeneity). Within-category homogeneity is called an *auxiliary assumption* (Waller & Meehl, 1998, p. 17), because it is an ideal situation; nevertheless, simulation studies have shown that violations of

this assumption can occur without detrimental effects for the approach (Beauchaine & Beauchaine, 2002; Waller & Meehl, 1998). Moderate correlations within categories (i.e., moderate within-category heterogeneity) do not hamper the application and power of the taxometric approach (Meehl & Golden, 1982). Large correlations within categories (i.e., large within-category heterogeneity) can be handled using an extension of the MAXCOV approach (Meehl, 1995). Still, the basic idea is that categories are relatively homogeneous by comparison with the between-category differences. Concentrating on relative homogeneity as the concept of category-likeness is reasonable; homogeneity is also a basic assumption in the latent class model. Two other features of the taxometric approach are (a) its limitation to applications in which only two categories are investigated (Beauchaine & Beauchaine, 2002) and (b) that it is less appropriate for binary data (Miller, 1996; J. Ruscio, 2000). According to Haslam and Kim (2002), about half of the studies to date make use of dichotomous indicators; they concluded that taxometric methods are valid for dichotomous indicators as well but cautioned that large sample sizes are required. They also recommended that "researchers should use continuous indicators whenever possible, but not shrink from using dichotomous indicators when there is no alternative" (p. 306). This recommendation contrasts with the fact that Dimcat applies equally well to dichotomous and polytomous indicators.

In sum, various methods relate to our approach, and each stresses one aspect of category-like structure. They are based on an underlying concept of category-likeness as showing abrupt between-category differences (multimodality), discrimination equivalence (factorial equivalence in its limited sense), or relative homogeneity of latent categories along a latent dimension (MAXCOV). Implicit in all of these approaches is the assumption of a mainly monothetic definition of category-likeness (but see the earlier quotation from Waller and Meehl, 1998, p. 9). The difference with our approach is that we explicitly include all of these aspects of category-likeness within a broader framework, one for manifest categories (in contrast with taxometrics). A category can be category-like in different ways, and a dimension can also be dimension-like in different ways. In this polythetic definition of category-likeness, being category-like is both complex and a matter of degree.

Three Applications

In this section, we describe applications to (a) personality disorders, (b) attitudes toward capital punishment, and (c) stages of cognitive development. In all three applications, manifest categories were defined either on the basis of expert judgment (by clinicians for personality disorders, by respondents for attitudes) or on the basis of segmentation (for developmental stages).

Application 1: Dramatic, Erratic Personality Disorders

In psychiatry, one used to think of disorders as categories of persons with a typical pattern of symptoms, called a *syndrome.* The categorical view, however, came under attack, especially with regard to personality disorders (e.g., Livesley et al., 1994; Widiger, 1992). First, patients within a category showed heterogeneous symptoms. Second, disorders seemed to come in degrees, both within the category and in comparison with the absence of the disorder. A twofold reaction to these findings has included (a) revision of the diagnostic system and (b) research on the dimension–category issue.

Psychiatric diagnosis has come to rely primarily on matching of features on a list provided by the *DSM–IV.* The syndromes defined by such features are supposed to be atheoretical and purely descriptive. The categories of the *DSM–IV* are not categories in the classical sense, defined by singly necessary and jointly sufficient criteria; rather, they are more akin to prototypes, because they are defined by showing a certain number of features from a list, with each feature typically being equally weighted.

Researchers have shown how a prototype approach can be applied directly to the classification of psychopathology. The prototype view has been contrasted with the classical view of psychiatric diagnosis (Cantor, Smith, French, & Mezzich, 1980). For example, a prototype approach has been applied to the classification of BPD (Clarkin, Widiger, Frances, Hurt, & Gilmore, 1983). Indeed, the concept of mental disorder itself has been speculated to constitute a prototype (Lilienfeld & Marino, 1995).

The *DSM–IV* reflects a revision such that diagnosis is based on showing a critical number of symptoms from a list, independently of the specific symptoms shown. This approach allows for heterogeneous symptom patterns, on the condition that they come from the list of symptoms associated with the disorder. The *DSM–IV* authors did not go so far as to reject the idea of categories altogether. One may wonder what is the basis for resistance against giving up the notion of personality disorder categories altogether. The resistance may be inspired by a cognitive bias toward thinking in categories, which may lead some to feel that categories of personality disorders tally with their experience of reality. Social psychology has a tradition of theories based on the assumption that people tend to categorize other people (e.g., Tajfel, 1981), and this is also the view in cognitive psychology (e.g., Smith & Medin, 1981). This argument has been invoked by Beauchaine and Waters (2003) to cast doubt on methods that are based on ratings.

The issue of whether disorders are category-like or dimension-like has become a topic of research and debate. A large majority of studies reject the categorical view in favor of the dimensional view. Three main empirical arguments have been presented for the dimensional view of personality disorders. First, personality disorders do not show bimodality (e.g., Kass et al., 1985; Nestadt et al. 1991; Zimmerman & Coryell, 1990). Second, personality disorders show factorial equivalence in its limited sense (e.g., Livesley et al., 1992). Third, personality disorders do not show relative homogeneity as derived from the MAXCOV procedure (e.g., Trull et al., 1990).

The issue we are raising is deeper than the formal issue of whether one should treat personality disorders as category-like, dimension-like, or some combination. If substantial qualitative differences exist, then the meaning of a symptom differs depending on the group to which a person belongs. Thus, the issue has consequences for both the theory and assessment of symptoms and syndromes of psychopathology. A consequence for diagnostic purposes is that a simple score based on symptoms, such as a sum score, can no longer be compared from one group to another.

In the present study (on the basis of Maesschalck's, 1998, study), we focused on BPD as compared with two other person-

ality disorders of Cluster B (the dramatic, erratic cluster): HPD and APD. These three disorders were compared with respect to the *DSM–IV* symptoms for BPD. In this connection, we noted above that one aspect of the dimension–category issue is relativity to the groups compared.

Some words of caution are needed for one to see the study in the correct light. First, we used a particular selection of indicators, and the results may depend on the indicators considered. This is a basic feature of our approach and of all other current approaches. This is what we meant by deeming our approach *relational.* Second, we used ratings by clinicians. Ratings do not necessarily reflect the truth. Because we relied on ratings and classifications, one should be aware that the ratings and the diagnosis are not perfectly reliable, although they come from experts. The less-than-perfect reliability may lead to a larger heterogeneity within the categories. We do not intend to investigate the true disorder categories, however, but rather the assigned disorder categories (i.e., manifest categories). This is of interest because most category-like variables in psychology are manifest. As a consequence, our conclusions must be seen as being based on categories that are assigned by experts, and we cannot claim more than cognitive relevance of the results. In other words, there is no "gold standard" for diagnosing psychopathology (Sher & Trull, 1996). This brings us to the cognitive approach to categories that we described in the introduction. Third, the manifest categories are not mutually exclusive. In psychopathology, overlap is called *comorbidity.* In our study we did not include patients with a multiple diagnosis for several reasons: (a) Overlap creates new manifest categories, so-called conjunctive categories (composed of patients with multiple diagnoses), and their structure is quite complex (Storms, De Boeck, Hampton, & Van Mechelen, 1999), so that it seems reasonable to start with pure manifest categories; (b) the inclusion of multiple diagnoses may confound the results in a way that cannot be detected, because there are not enough cases of each different multiple diagnosis; (c) it is not without importance to investigate the latent structure of pure manifest categories, because they reflect the disorder in an unconfounded way. As a consequence, we are not able to generalize our results to the whole categories of the three diagnoses, but we believe that just as in a psychological experiment it may be of interest to create pure conditions.

Method

Participants. The sample was composed of 370 Dutch-speaking Belgians from 30 inpatient, outpatient, and prison facilities: 122 were diagnosed with BPD, 123 were diagnosed with HPD, and 125 were diagnosed with APD. The BPD group was 74% women and 26% men, the HPD group 77% women and 23% men, and the APD group was 14% women and 86% men. With regard to marital status, the BPD group included 65% single participants and 35% married participants, the HPD group included 43% single participants and 57% married participants, and the APD group included 54% single participants and 46% married participants.

Manifest categories. Axis I and Axis II diagnoses were made in the three weeks after first admission or consultation by one or more diagnosticians, usually including a senior psychiatrist. These diagnoses, which defined the manifest categories, were instances of expert judgment.

Indicators. Each patient was also rated by a clinician other than those on the initial diagnostic team on a list of nine *DSM–IV* symptoms of BPD. The rating clinicians were unaware of the original diagnosis. Note that this methodological feature of the study favors its objectivity, but at the same time makes it less relevant from a cognitive perspective. To draw conclusions of a cognitive kind, one would prefer that the same persons rate the indicators and do the categorization. Symptom ratings were based on information from charts, staff meetings, and contacts between the clinician and the patient. The symptoms were presented in a random order to be judged on a 4-point scale from 0 (*least severe*) to 3 (*most severe*). In the instructions, Scale Points 0 and 1 were defined as nonpathological, whereas Scale Points 2 and 3 were defined as pathological. Responses were later dichotomized, such that 0 and 1 were recoded as 0 (*less severe, nonpathological*), and 2 and 3 were recoded as 1 (*more severe, pathological*).

Analyses. The full modeling approach was followed as explained earlier, making use of SAS PROC NLMIXED for the dichotomized data, as explained in the Appendix. The BDP group was used as a reference category. The locations and discriminations in the other two categories were expressed as deviations from those in the BPD category. To test absolute goodness of fit, we used a bootstrap approach (Efron & Tibshirani, 1993). One of the aspects investigated is how well the correlations between indicators within each group could be explained from the model. Because we used unidimensional models within each diagnostic group, this bootstrap of correlations is also a test on the unidimensionality of the heterogeneous within-category structure. As mentioned earlier, Sanislow et al. (2002) presented a multidimensional model (but with extremely high correlations among the dimensions), so we wanted to ensure that we did not have to expand our model to be multidimensional as well (within each of the manifest categories). Note that it is possible to find unidimensionality within manifest categories, although the single dimension is different depending on the manifest category, implying that for the total group the model is multidimensional. When the persons belong to different manifest categories and a joint analysis is performed, one can conclude that the structure is multidimensional, whereas in fact it is unidimensional within each manifest category. Such a result would be in agreement with a Type 1 structure.

Results

We performed a group-wise principal-components analysis (Kiers, 1990) to explore the structure of the symptoms. The percentage of variance explained by a principal-components analysis within each group (on the basis of the 4-point scale ratings) was about as high as when a common solution was imposed on all three diagnostic groups (about 40%). Two of the nine BPD symptoms did not reach a loading of .30 on the BPD component and were therefore removed from further analyses. These symptoms were inappropriate anger and impulsivity in two areas. These symptoms were omitted from further analyses, so that seven symptoms remained. The symptoms were removed not because of group differences in discrimination but because of overall low discrimination. The two eliminated symptoms did not belong to a factor in the study by Sanislow et al. (2002), so the kind of multidimensionality found in that study cannot explain the poor results for the two symptoms. Other symptoms belonging to the same factor had rather large factor loadings in our study.

Following the strategy presented in Figure 3 and explained in the *Modeling* section, we began by investigating the nature of the between-category differences, on the basis of three models. Using a likelihood ratio test, it was found that the goodness of fit of the QUAL2–HET model was not statistically significantly worse than that of the QUAL1&2–HET model, $\chi^2(12) = 14.3$, $p > .10$. This means we can assume discrimination equivalence. When the QUANT–HET model was compared with the QUAL2–HET model, however, it turned out that its goodness of fit was worse, $\chi^2(12) = 56.5$, $p < .001$, which also implies that its goodness of fit was worse than that of the QUAL1&2–HET model. Therefore,

we cannot conclude that we have location equivalence—there seemed to be qualitative differences in terms of locations between the manifest categories.

To identify the location differences, we inspected the deviations of the locations in the HPD and APD groups from the BPD group using the reparameterization with α_i as a multiplication factor not just for theta but for the whole logit of item i and with effect coding for the location deviance parameters (see the Appendix for this reparameterization). Several of the deviation parameter estimates were statistically significant. In the HPD group, this was the case for the symptoms affective instability and avoidance of abandonment, with estimates of -0.747, $t(369) = 2.12$, $p < .05$, and -1.121, $t(369) = 2.87$, $p < .01$, respectively. In the APD group, a significant location deviance was found for chronic feelings of emptiness, (-0.833), $t(369) = 2.38$, $p < .05$, but also a rather large but nonsignificant deviation was found (-0.599) for avoidance of abandonment. For all four deviations, this implies that the corresponding symptoms were located lower on the dimension in the HPD group and/or in the APD group, and were thus more common. When for these four location differences one common saltus parameter was used and all other locations were assumed to be equal over the three groups, the resulting saltus model was not significantly worse than the QUAL2–HET model, $\chi^2(11) = 14.4$, $p > .10$, implying that it is sufficient to limit the location differences to this one saltus parameter and to two symptoms in each manifest category. This result is very similar to the one we obtained without the reparameterization, where (a) only one fewer symptom was given a saltus parameter (chronic feelings of emptiness, in the APD category) and (b) the difference in goodness of fit with the QUAL2–HET model was slightly larger and significant. We also compared the models on the AIC and BIC criteria: The lower the values, the better the model. The AIC value of the saltus model was slightly lower than that of the QUAL2–HET model (2,921.1 vs. 2,927.7), and its BIC value was clearly lower (2,995.4 vs. 3,045.1), so that it can be considered a good approximation. We should also mention that the AIC and BIC values of the QUAL1&2–HET model were 2,937.4 and 3,101.8, respectively, both higher than the corresponding values of the QUAL2–HET and QUAN–HET models.

When the model is further restricted to have zero variance within the three groups, the goodness of fit is dramatically lower following the likelihood ratio test, which is conservative given the boundary value of the null hypothesis, $\chi^2(2) = 180$, $p < .001$. Each of the variance estimates is highly significant in the QUAL2–HET model using a Wald test (which is also conservative in this case). Therefore we must conclude that the diagnostic groups were heterogeneous. This was corroborated by a statistically significant Cronbach's alpha in the three groups: .49 for BPD, .61 for HPD, and .67 for APD (all ps $< .01$). Taking together the conclusions regarding the vertical and the horizontal axes, we end up with a Type 1 structure: between-category qualitative differences and within-category heterogeneity. A reasonably good saltus model was found, so that the qualitative differences can be considered rather simple.

We now further explore the model that came out as the best, QUAL2–HET—a model with discrimination equivalence but not with location equivalence. This model implies a 2PL model within each diagnosis with equal discriminations between diagnoses. To test this model, we applied a bootstrap methodology. Starting from the parameter estimates, we generated 2,000 new data sets, and in each of these data sets the following statistics were derived: Pearson correlations (phis) between the indicators within each diagnostic group (yielding 21×3 correlations) and differences in assigned symptom proportions for the HPD and APD groups in comparison with the BPD group as the reference group (7×2 differences). Of the 63 correlations only 2 fell outside the bootstrap-based .01 confidence interval, and 3 more fell outside the corresponding .05 confidence interval. This is a remarkably good result, from which it can be concluded that the model and also its unidimensionality within groups should not be rejected. The result was even better where the proportion differences were concerned. All 14 differences fell right in the middle of the confidence interval, implying that the model captured the location differences very well. On the basis of this bootstrap result, we can accept the QUAL2–HET model.

Apart from the crucial aspects of this model to decide on the type of latent structure (in this case Type 1), some other aspects of the model are of interest. First, the variances in the three groups differed. The variance in the BPD groups was fixed to 1.000 as an identification restriction, and the estimates in the other two groups were 1.292 (HPD) and 2.288 (APD). These differences were in agreement with the size order of the internal-consistency coefficients that were reported earlier. Larger variance typically means larger consistency. Second, HPD and APD were less borderline than was BPD. The difference of HPD from BPD was -1.452, and the difference of APD from BPD was -2.748. Both difference estimates (on the theta scale) were statistically significantly different from zero ($p < .001$), meaning that overall group effects were statistically significant. The most borderline group was BPD, as expected, followed by HPD and APD.

Similar studies were conducted on the diagnoses of HPD and APD, using histrionic and antisocial symptom lists from the *DSM–IV,* respectively (Maesschalck, 1998). For HPD, the result was similar, in that only simple qualitative differences in location were found. For APD, however, the qualitative differences could not be reduced to a few saltus parameters; the pattern of APD indicator values was quite different among the diagnostic groups, as shown in the left panel of Figure 2.

Discussion

Strictly speaking, with respect to BPD symptoms, there is evidence for qualitative differences between the three groups. These differences can be attributed to a few symptoms. Only two symptoms in each manifest category showed a statistically significant location deviation, and a saltus model with only one saltus parameter yielded a very good approximation. Taken together, there seems to be evidence for the following: (a) affective instability is relatively more common in the HPD group than in the other two groups, (b) chronic feelings of emptiness are relatively more common in the APD group than in the other two groups, and (c) avoidance of abandonment is relatively more common in both the HPD and APD groups than it is in the BPD group. This summary is based on the statistically significant deviations and on the saltus model. Note that the result for avoidance of abandonment could be attributed to deficiencies in the indicator rather than to qualitative differences in the diagnostic groups. Specifically, most of the APD

patients were prisoners, so they were likely to show this symptom by reason of their isolation in prison rather than their personality disorder. This result illustrates the general point that there are two kinds of qualitative differences: those that indicate true qualitative differences due to the manifest categories and those that indicate differences due to irrelevant reasons. With regard to patients in prison, the abandonment symptom was a theoretically poor indicator; thus it is not surprising that it also showed a location difference. The best alternative in such a case is probably to remove the indicator from consideration. For the same symptom, however, an even stronger location difference was also found with the HPD group. In sum, the two diagnoses show simple and rather weak qualitative differences of a kind that would not have been detected with a simple test for factorial equivalence in its limited sense.

This conclusion cannot be taken as an absolute, because of the restrictions we mentioned earlier. The data concern only a limited number of indicators, although very important ones, and they are based on ratings by clinicians. Because of the latter, our conclusion must primarily relate to the dimension-like versus category-like nature of judgments made by clinicians. As such, the results can also be looked upon from the cognitive perspective on categories. The clinicians' category of BPD (independently of whether it reflects the true state of affairs) is a manifest category with a latent continuum, with some BPD members being better members of the category than others. The result may have been cognitively induced, although the experts who rated the indicators were different from those who made the diagnosis. The structure within the manifest category is unidimensional: the stochastic variant of what has been called a *triangular structure.* The HPD and APD patients not only are less borderline but also show some slight qualitative differences, enough to conclude that BPD is category-like in at least one respect: that of qualitative between-category differences.

Our findings regarding BPD may not generalize to other categories of personality disorders as may be derived from taxometric studies (although they follow a different approach). For example, studies have found taxometric evidence for the taxonic nature of schizotypy (Golden & Meehl, 1979; Korfine & Lenzenweger, 1995; Lenzenweger, 1999; Lenzenweger & Korfine, 1992) and of APD (Skilling, Quincey, & Craig, 2001), whereas some evidence favors a dimension-like structure for BPD (Rothschild et al., 2003). This shows that being category-like may depend on the personality disorder, which was also the case for the data we used (Maesschalck, 1998) showing that APD is more category-like than are BPD and HPD.

The phenomena we identified at the latent level can be considered endophenotypes. These refer to the phenotype but go deeper than the manifest indicators. When category-like, endophenotypes comprise natural kinds, nonarbitrary discontinuities; when dimension-like, they comprise equally nonarbitrary continuities. Haslam (2002) noted,

> Of course, a discrete psychopathological kind might arise out of an essence-like cause such as a genetic abnormality (e.g., Down's syndrome) or germ (e.g., general paresis). However, other nonessentialist models are also possible, for example developmental polarization, nonlinear interactions of vulnerability factors (e.g., emergenesis), and threshold effects. (¶ 13)

A continuous endophenotype, by contrast, is likely to result from divergent causes, such as polygenic influences, idiosyncratic environments, and "bad luck" (cf. Meehl, 1978). When an essence-like cause becomes known, an endophenotype becomes a closed concept, but, contrary to the essentialist beliefs of most laypersons (Haslam & Ernst, 2002), most endophenotypes in psychopathology (including category-like ones) have no essence-like cause and thus remain open concepts.

It is somewhat surprising that the identification of endophenotypes has not always been the primary concern in the classification of psychopathology. Instead, the operational approach was espoused in order to increase interjudge reliability. The consequent increase in reliability was purchased at the price of a decreased theoretical basis (e.g., Carson, 1991) and, more formally, a lack of interest in the latent structure (Acton & Zodda, in press). This contrasts with the explanatory approach in cognitive psychology, discussed above (e.g., Murphy & Medin, 1985), in which the glue that ties concepts together is a theory-based understanding of the world (e.g., Kim & Ahn, 2002). Although we did not investigate the theoretical basis of the diagnostic categories, we assessed the validity of several latent structure models of BPD. As far as the endophenotypes are concerned, we were able to find out what the BPD endophenotype is—not all aspects of it, but those related to the *DSM–IV* borderline indicators. Now the task remains of incorporating the open concept of the BPD endophenotype into a larger nomological network, including theories of its etiology, course, and treatment.

Beyond the question of the category-like versus dimension-like latent structure of psychiatric diagnoses, at least three controversial issues within psychopathology and treatment research could be addressed using Dimcat. The first issue is whether putatively distinct disorders are not really identical. Consider several examples on the border between Axis I and Axis II: avoidant personality disorder and social phobia, schizotypal personality disorder and schizophrenia, BPD and mood disorders, APD and substance use disorders, and depressive personality disorder and dysthymia (Endler & Kocovski, 2002; Widiger & Shea, 1991). Frances, Widiger, and Fyer (1990) noted,

> It is rarely clear, when a given symptom serves as a defining feature of two different categories, whether the resulting overlap between them reflects the true state of the relationship or is an unnecessary artifact based on the choice of the identical definitional items in both sets. (p. 47)

From the perspective of Dimcat, this question can be answered rather straightforwardly. A combined symptom list could be taken as a list of indicators. Whether the symptoms overlap does not matter. If the disorders were qualitatively distinct, then they would obviously not be identical. If the disorders were only quantitatively distinct, then they would be identical if the difference between the distributions was of a magnitude considered pragmatically negligible.

The second issue is whether a psychiatric diagnosis can be adequately assessed by a self-report inventory. This issue has been debated with respect to using students who score high on the Beck Depression Inventory as "analogs" of patients diagnosed with major depression (e.g., Coyne, 1994; Flett, Vredenburg, & Krames, 1997; A. M. Ruscio & Ruscio, 2002; Vredenburg, Flett, & Krames, 1993). From the perspective of Dimcat, this question

can also be answered rather straightforwardly, taking the items in the inventory as indicators. If the diagnosis was qualitatively distinct from its absence, then the self-report inventory would not be an adequate representation of the diagnosis—it would mean that the inventory was measuring qualitatively distinct phenomena for persons with and without the diagnosis. If the diagnosis was only quantitatively distinct from its absence, however, then the latent dimension defined by the self-report inventory fulfills a necessary condition to be an adequate representation of the diagnosis.

The third issue is whether a stepped-care approach to treatment is appropriate. In a stepped-care approach, treatments are tailored to the level of severity of the disorder. Such an approach presupposes heterogeneity within the category of persons with a diagnosis (e.g., Acton, Kunz, Wilson, & Hall, in press). For example, a stepped-care approach for the treatment for *DSM–IV* nicotine dependence might be recommended, such that a stop-smoking pamphlet or telephone quitline might help some smokers, whereas other smokers might require an antidepressant or extensive cognitive-behavioral treatment. This would make sense only if this manifest category were heterogeneous, with some nicotine dependent smokers higher on nicotine dependence than others.

Application 2: Attitudes Toward Capital Punishment

Capital punishment is a controversial issue. In Belgium, capital punishment was legal until 1996, but it had not been practiced since 1950. In 1996, Belgians voted to ban capital punishment at a time when it was no longer a real issue. Later that year a man was accused of kidnapping, raping, and murdering several girls between the ages of 8 and 17. This case was in the news for over a year and, not surprisingly, affected many people's opinions of capital punishment. The legalization of capital punishment again became a topic of heavy discussion and a source of controversy. At the time we conducted the present study, there seemed to be two clear-cut public opinions: one in favor of capital punishment and one opposed.

We studied attitudes toward different types of crimes varying in the following characteristics: murder or other crimes, sexual or nonsexual crimes, and child or adult victim. A group of respondents was interviewed and asked whether, in their opinions, persons who committed the kind of crime in question should be considered for capital punishment if it were legal.

Our first interest was whether the attitudes were qualitatively distinct. This kind of question is not uncommon for attitude research. Eagly and Chaiken (1993), for example, asked whether the relation between liberalism and conservatism, which might seem opposite poles of a single dimension, was actually more complex. One explanation for the latter structure would be that the two groups differ in the values considered relevant to an issue. In the present context, these may concern the unconditional value of human life, the acceptability of revenge, and the seriousness of a crime. The criteria for seriousness of a crime may include taking someone else's life, sexual abuse, and vulnerability of the victim. Differences in these criteria should result in a qualitatively different scale for seriousness of a crime between groups in favor of and opposed to capital punishment.

Our second interest was whether the attitudes were heterogeneous. Only if the attitude groups were heterogeneous could within-category person differences be observed.

Our third interest was in the capacity of our approach to differentiate between a purely manifest continuum versus a latent continuum. The reason is that in this application it would not be a surprise if there were two clear-cut homogeneous attitudes in the latent structure. In the case of latent homogeneity, one can be misled by the heterogeneity that would show up not only in the sum scores but also in the estimates of individual thetas. The crucial test, however, is not in the sum score or theta estimate distributions that would result but in (a) the likelihood ratio test to compare a heterogeneous with a homogeneous model and (b) the test of each variances' difference from zero. Therefore, we set up a simulation study to investigate whether we could differentiate between categories being heterogeneous or homogeneous in their latent structure, notwithstanding the expected heterogeneity in the sum scores and the estimated thetas. This issue is also important because Haertel (1990) showed that the 2PL model can be approached quite well with a latent class model.

Our fourth interest was related to the study of cognitive categories. Because the data in this application were self-rating data, and because the rating of the indicators and the classification were both made by the same respondents, a cognitive approach to the categories seemed relevant. This offered us an opportunity to test the generalized context model, a model for how people decide on a category, because it focuses on classification into two categories and because the respondents both classified themselves in two categories and made the indicator ratings. Assume that the respondents decided on whether they were in favor of or against the legalization of capital punishment from what they heard from others. For example, they heard what other people said about various crimes and how the criminals should be treated. These other people can be considered the exemplars of the learning set, before the respondents decided on the classification of their own opinion. The alternative to the exemplar theory is that the self-classification in the two legalization opinions is based on two prototypes.

Method

Participants. In several small towns along the Belgian coast, 300 adults (50% women, 50% men) were interviewed in 1998 about various types of crime. At that time, the above-mentioned case of child abuse was still very alive in the minds of Belgian people, as indicated by the attention the case received in the media. In response to a single question at the end of the interview, 202 respondents were in favor of legalizing capital punishment, and 98 were against.

Indicators. The interview consisted of 10 questions, 9 of which referred to the following crimes, in this order: (a) serial murder, (b) murder of one's whole family, (c) murder of a family member, (d) sex murder of an adult, (e) sex murder of a child, (f) robbery with murder of an adult, (g) robbery with murder of a child, (h) rape of an adult, and (i) rape of a child. For each crime, the question was whether the respondent would consider capital punishment appropriate if it were legal ("yes" or "no"). The 10th question was whether the respondent was for or against the legalization of capital punishment.

Manifest categories. Two manifest categories of attitudes were distinguished on the basis of the 10th question: one in favor of legalization and one against legalization. These categories were based on expert judgment, with respondents considered experts on their own attitudes.

Analyses. The main part of the analyses were again based on Dimcat. Because we experienced estimation problems with the more complex models, most likely due to the manifest distribution of the data, we based

part of the analysis on a CML approach using the OPLM computer program (Verhelst et al., 1994).

Two additional analyses were run. First, for the simulation study we used two homogeneous categories with nine indicators and the following betas: .20, .20, .30, .15, .35, .30, .20, .35, and .25. There was a .40 difference on the theta-scale between the two categories (one theta equals 0.00 and the other equals .40). Ten data sets were generated with these parameters and 300 persons in each category. The data were analyzed with the QUAN–HET model (with equal and unequal discriminations) and the QUAN–HOM model.

Second, to analyze the data following the generalized context model, we made the (arbitrary) choice (a) to select the response patterns of 40 randomly sampled respondents of each group as the learning stimuli and (b) to select the remaining response patterns as the test stimuli. This is as if the respondents first had been informed about 40 people's opinions (through daily life discussions) before they decided on their own attitude category (in favor of or against legalization) on the basis of what they think of how the criminals should be treated. The procedure was repeated five times, each time with a randomly sampled learning subset from each group and with the remaining respondents as the subset of test stimuli. The prototypes for the two categories were defined on an a priori basis. As the prototype for the prolegalization category, we took the overall 1-pattern for all indicators, and as the prototype for the antilegalization category, we took the overall 0-pattern for all indicators (the complement of the first prototype). To compare the prototype model to the exemplar-based model, we used the same five sets of test stimuli for the two models.

For both models, we used the nine binary indicators as nine binary features or dimensions. The maximum-likelihood-based analysis was performed with two different similarity functions (one with an exponential decay [$q = 1$], another with a Gaussian decay [$q = 2$]) and with a city-block metric (because of the binary features and a better goodness of fit than the Euclidian metric). Eleven parameters were estimated for both models: c (an overall scaling parameter—the higher its value, the larger the weight of close similarities), b (response bias toward the category in favor of legalization), and nine indicator weights (eight of which were free parameters, given that their sum is 1).

Results

We again used the sequential modeling strategy explained earlier. The first model to be tested, however, the QUAL1&2–HET model, yielded convergence problems and extreme parameter estimates. This was also true for simpler models with estimated discriminations and different variances depending on the group. A possible reason for the problems was the distribution of the persons. The frequencies of the 10 possible sum scores were as follows: 63, 6, 6, 6, 16, 15, 27, 32, 39, and 90. Of 300 respondents, 63 would not consider capital punishment for any of the nine crimes, and 90 would consider it for all nine crimes. The proportions of "yes" responses for the nine crimes are given in Table 1. Because of this unusual distribution, we decided to shift to a CML approach, because it is free of distribution assumptions. The one-parameter logistic model (OPLM) is based on conditioning on sufficient person statistics, and it is therefore saturated with respect to the person distribution (de Leeuw & Verhelst, 1986). Using the discrimination values suggested by the OPLM module for both attitude groups (the same for both), we fit a model with discrimination equivalence and location equivalence to the data with the OPLM program (Verhelst et al., 1994). The model fit the data quite well when tested with a Pearson-chi-square-based test statistic: the R1c described by Glas (1988). The R1c value was 15.75 ($df = 17$, $p > .10$). Thus, we can conclude that a Type 2 structure had a reasonable goodness of fit.

Next we estimated a QUAN–HET model with SAS PROC NLMIXED with the same fixed discrimination values (see Table 1) and also with location equivalence. The resulting deviance was 1,630.9, and the corresponding AIC and BIC values were 1,654.9 and 1,699.4, respectively. The deviance of this model was only slightly higher than that of the corresponding CML model (1,630.9 versus 1,625.9), so that the difference in distribution between the two approaches did not seem to play an important role in the goodness of fit for the QUAN–HET model. Note that the discriminations of the individual indicators cannot be estimated very reliably when the sample size is rather small. Because we will not interpret these individual discriminations, and because of the previous result, we constrained all discriminations to be equal within and between categories. Analogously, the theta estimates for individual persons would perhaps not be very reliable when only nine indicators are used, but again we concentrate on overall features, such as the parameters of the theta distribution(s). The result of the QUAN–HET model with equal discriminations was a deviance of 1,582.0, with corresponding AIC and BIC values of 1,606.0 and 1,650.4, respectively. From this result it seemed that equal discriminations were a good option when a normal distribution was assumed. Assuming equal discriminations for all indicators, we estimated a QUAL2–HET model in the next step, which is actually a step back in the order of testing. The resulting deviance was 1,576.7. Based on a likelihood ratio test, this is not statistically significantly lower than the deviance of the QUAN–HET model with equal discriminations, $\chi^2(8) = 5.30$, $p > .10$. Accepting location differences between the two groups did not seem to pay off. Therefore, we continued with the QUAN–HET model with equal discriminations for all indicators as the reference model.

We tested this model against the QUAN–HOM model to make a choice along the vertical axis in Figure 1. The resulting deviance was 2,133.1, and the corresponding conservative likelihood ratio test was statistically significant, $\chi^2(2) = 551.1$, $p < .001$. The conclusion must be that the QUAN–HET model was the better one and that the groups were heterogeneous.

From an inspection of the QUAN–HET parameter estimates, note that the two attitude groups seemed to differ in attitude level as well as in heterogeneity. When reporting the estimates, we mention the standard errors in parentheses. The estimate of the group effect on the latent continuum was −8.847 (0.847), which was statistically significant. The group that was against capital punishment was located much lower on the attitude continuum

Table 1
Proportions in Favor of Capital Punishment ("Yes" Responses) for the Nine Crimes

Crime	Proportion "yes"	Fixed degree of discrimination
Serial murder	.77	2
Sex murder of a child	.75	3
Murder of one's own family	.67	2
Sex murder of an adult	.67	2
Robbery with murder of a child	.63	3
Rape of a child	.63	1
Robbery with murder of an adult	.60	3
Murder of a family member	.51	3
Rape of an adult	.36	1

than was the group that was for capital punishment. The variance of the two attitude groups was quite different: $\sigma^2_{pro} = 4.227$ (.842), and $\sigma^2_{anti} = 17.391$ (3.404). Both estimates were statistically significantly different from zero using the conservative Wald test for variances. This confirmed the earlier conclusion that the groups were heterogeneous. The difference between the two variances was also estimated (in a separate run). The result was 13.164 (3.444), which was statistically significant using a Wald test. The latent structure for the two groups seemed to be one with a relatively homogeneous group in favor that is rather far above a much more heterogeneous group against.

To provide a better view on the latent distribution, in Figure 4 we show a histogram of that distribution based on the estimated distribution parameters. Because the size of the groups clearly differs and may have a misleading visual effect, we constructed the histogram for groups with equal size (both $n = 202$, which is the size of the largest group). The distribution was clearly bimodal and corroborates the bimodality of sum scores. Because there was also a clear difference between the means of the two groups, the quantitative difference between the two attitude groups may be considered to be abrupt.

The data of the simulation study were first analyzed using equal discriminations (as they were generated and in conformity with the results on capital punishment) with the QUAN–HET and QUAN–HOM models. The results show smooth histograms without any gap for the sum scores as well as for the individual theta estimates from the QUAN–HET model with no category main effect. One might be misled to conclude that the latent structure of the categories is heterogeneous. Using a likelihood ratio test, however, the QUAN–HOM model was never rejected against the QUAN–HET model independently of the estimated category effect (all ps $> .10$ and differences in the deviance statistic that are smaller than 1.5), and in none of the 10 data sets was the variance statistically significantly different from zero (all ps $> .10$). Similar results were obtained with the 2PL model. This result also means that we can differentiate between a model with homogeneous classes and the 2PL (with or without a category main effect), so that our concern based on Haertel's (1990) study is met. The result of this small simulation study is reassuring for our approach. It shows how one can be misled by apparent heterogeneity in the sum scores and in the individual theta estimates if one does not use statistical tests for features of the latent structure. Consequently, the bimodal distribution in Figure 4 should be seen in the light of the statistical tests.

As for testing the exemplar model and the prototype model with $q = 1$, the means of the log likelihoods were 78.6 and 78.9, respectively, and for $q = 2$ the corresponding values were 76.5 and 78.9, respectively. (The value of q will not make a difference for the prototype model because of the way the prototypes were defined.) This means that the two models performed about equally well. For the prototype model, the c estimate varied between 1.98 and 9.54, whereas the corresponding values for the exemplar model were more extreme—from 6.43 to 15.97 for $q = 2$, and even more extreme for $q = 1$. High values of c mean that close similarities weighed much more heavily in determining the classification decision. The b estimates were found to be in line with the fact that the group in favor was larger than the group against. Finally, the weights were more stable (over the five runs) for the

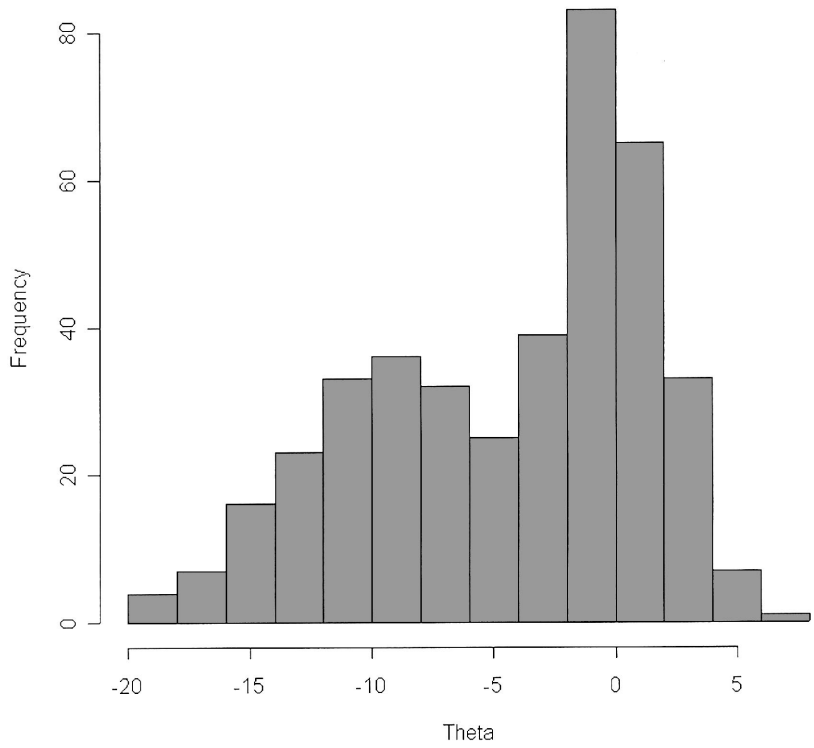

Figure 4. Empirical Bayes joint distribution of two manifest categories (with number of observations on the ordinate).

prototype model than for the exemplar model. The highest average indicator weights in the prototype model were found for serial murder (.349), murder of one's whole family (.239), and rape of a child (.132; the same for the two values of q).

Discussion

A latent structure with heterogeneous quantitative and abrupt differences between attitude groups appears to describe Belgian attitudes toward capital punishment. No evidence was found for location differences, and a CML model with location equivalence also seemed to fit the data in an absolute sense. Thus, the two attitude groups can be considered to be located along the same latent dimension, although at a different point on that dimension, with a gap in between and with a different variance.

If one were to invoke the bimodal distributions as evidence for the existence of a latent categorical structure, then one should realize that the bimodality is a relative criterion, namely the size of the main effect of the group factor. All other aspects of the latent structure are dimension-like. Because there is no definitive way to tell how large the absolute difference should be, nor how large Cohen's d should be, and because the bimodality follows from the size of Cohen's d, the bimodality is at best a relative criterion. That the two categories appear as heterogeneous is not an artifact and neither is it derived from the histogram in Figure 4. It is based instead on the result of a likelihood ratio test and a test of the variances. The discriminative power of our approach was corroborated through the results of a small simulation study.

The conclusion that the structure is dimension-like (apart from the abrupt difference) needs a word of caution. First, one can imagine that indicators could be used other than the nine we studied. For the personality disorder categories, the selection of indicators (the symptoms) had a strong basis in the *DSM–IV*. For the attitudes toward capital punishment, the choice was less evident. Second, ratings were again used, but they were self-ratings instead of ratings by experts. Given that the indicator ratings and the classifications were made by the same persons, the conclusions may reflect the cognitive construction of attitudes by the respondents.

As to the relevance of the cognitive models for our data, there is no way to compare the goodness of fit of the exemplar-based and prototype models with the nonlinear mixed models that we estimated. The purpose and the structure of the models are totally different. In Dimcat, the classification (the manifest category) is a predictor for the indicators, whereas in the cognitive models, the indicator data are the predictors for the classification. The structure of the cognitive models is also quite different—for example, because of the crucial role of similarities between exemplars or of exemplars with the prototype. There is no counterpart of this in the nonlinear mixed model family.

The fact that the performance of the prototype model is about as good as that of the exemplar-based model is remarkable. It would be of interest from a cognitive-psychological viewpoint to compare two types of categories, one with manifest heterogeneity but no internal structure and another with manifest heterogeneity and latent heterogeneity, to investigate whether the superiority of the exemplar-based model generalizes to dimension-like (heterogeneous) categories. As discussed earlier, within-category structure has been neglected thus far in the cognitive literature. Our results could inspire studies to investigate the effect of the within-category structure on the validity of the exemplar model and the prototype model.

Application 3: Stages of Cognitive Development

Several stage models of cognitive development have been formulated. The saltus model (Wilson, 1989) was developed to overcome the limitations of other stage models, described below.

First, the scalogram model (Guttman, 1944) has been applied to stage-like development (see Kofsky, 1966, for a critique). This model is deterministic, meaning that performance on different cognitive problems is perfectly determined by the stage reached. The model implies that the stages are homogeneous and linearly ordered.

Second, the multitask approach (K. W. Fischer, Pipp, & Bullock, 1984) was developed to relax the limitation that stages need to be homogeneous, in order to capture microsequences within the stages. K. W. Fischer et al. (1984) made an interesting distinction between first-order versus second-order discontinuity, a distinction similar to our distinction between quantitative versus qualitative differences. A first-order discontinuity is a sudden leap in performance (corresponding to quantitative differences on all relevant problems), equal for all problems, whereas a second-order discontinuity is a discordant leap (corresponding to qualitative differences), large for some problems but not for others. K. W. Fischer et al. (1984) accepted the probabilistic link between stages and solving problems but did not use the idea for formal modeling.

Third, the ordered latent class model (Croon, 1990) can be used to relax the deterministic nature of the model (and of the stages). It provides an explicit probabilistic link between stages and performance on problems. Within-stage homogeneity is still assumed, as in the scalogram model, albeit homogeneity of a stochastic kind. Although the classes (stages) are ordered, they can show qualitative differences, because problem locations can differ across classes. Indeed, the problem locations must meet certain inequality restrictions for the classes to be ordered (see also Hoijtink & Molenaar, 1997). The ordered latent class model is situated between Type 3 and Type 4 from Figure 1, but for latent categories.

In contrast with these three models, the saltus model combines a probabilistic view of stages, the assumption of within-stage heterogeneity, and the possibility of modeling certain between-stage qualitative differences. The saltus model has a special type of parameter to distinguish between first-order and second-order discontinuities, the δ-parameters. A $\delta_{kks'} \neq 0$ implies that for stage k in comparison with stage k', performance on a subset s of problems differs from performance on the complementary subset of problems. Differences of this kind are qualitative, because differences between problem locations are not equivalent across stages. When no saltus parameters are required (the saltus parameters are zero) and the stage main effects suffice, the discontinuities are of the first-order type and quantitative. For a first-order discontinuity to occur, the distance between groups of persons on the latent dimension (which is also the proficiency scale) must be large—for example, without overlap. In sum, the saltus model lacks the limitations of the previous models, and it allows for the distinction between two kinds of discontinuities. Furthermore, the saltus model is a particular specification of a Type 1 model from Figure 1.

Saltus parameters can capture how some problems become much easier relative to others as persons add to or reconceptualize their knowledge. Saltus parameters can also capture how some problems actually become harder as persons progress from an earlier stage to a more advanced stage, because they previously gave the correct answer but for the wrong reasons. There are two ways to apply the saltus model. One way (in which it was originally developed) is to assume that class membership is a latent variable estimated from the data—we will call this the latent saltus model (Mislevy & Wilson, 1996; Wilson, 1989). A second way is to assume that class membership is an observed variable that is given by, for example, segmentation or expert judgment—we call this the *manifest saltus model* (G. Fischer, 1992; Wilson, 1993). The assumption of manifest class membership makes estimation of the model simpler, and it may make interpretation more straightforward, but it also involves certain limitations (Wilson, 1993).

A Rule Assessment Hierarchy Approach

Siegler (1981) developed modified Piagetian problems to test the cognitive developmental theory rule assessment. The most important characteristic of the rule assessment approach is the specification of a series of increasingly powerful rules for solving problems. Following this theory, the behavior of a learner is dominated by the rule he or she is using at a particular level of development (a particular stage). The sequence of development through the rules is assumed to be fixed. The theory differs from a Piagetian approach in that (a) the rules do not need to be the same across concepts, and (b) the indicators are nonverbal choices to concrete problem-solving tasks.

Siegler (1981) investigated the rule assessment theory with three experimental problems involving proportionality: a balance-scale problem, a projection-of-shadows problem, and a probability problem. We concentrate on the balance-scale problem. Using problem analysis and by reference to previous empirical and theoretical work, Siegler posited a series of rules that children might use in tackling the problem. A child using Rule I will not consider the distances of the weights from the fulcrum; to such a child, only the amounts of the weights matter (weight is the dominant dimension). A child using Rule II will consider the distances of the weights from the fulcrum only when the weights are the same (distance is the subordinate dimension); otherwise the child will consider only the amounts of the weights. A child using Rule III is aware of his or her lack of understanding of the behavior of the balance scale when both weights and distances vary and will use a cognitive strategy such as guessing or taking cues from the experimenter. A child using Rule IV will compute torques on either side of the balance beam and choose accordingly; this computation can be executed either by actual calculation or by "eye."

To distinguish between persons at these four rule levels, Siegler (1981) designed six types of problems, of which we present three: dominant problems (D), with unequal values on the dominant dimension (weight) and equal values on the subordinate dimension (distance); subordinate problems (S), with equal values on the dominant dimension (weight) and unequal values on the subordinate dimension (distance); and conflict-equal problems (CE), with unequal values on both dimensions but with the two sides balanced (see Figure 5).

The six problem types yield different profiles for the four rules, and this difference was the basis for Siegler's classification. For the three kinds of problems we described, the differentiation is as follows. Rule I differentiates between D problems and S problems, because D problems can be solved when exclusively the dominant dimension is used, but S problems cannot. Rule II differentiates between D or S problems and CE problems, because taking the subordinate dimension into account in the case of equality on the first dimension helps a person solve S problems but not CE problems. Rule III differentiates in a similar way, except that a person will guess on CE problems. Finally, Rule IV also will lead a person to guess on CE problems, because the combination of distance and weight on both sides yields a tie. The three problem types considered here permit the distinction between adjacent rule levels: D versus S (Rule I vs. higher), and S versus CE (Rule II vs. higher). The three stages are differentiated on the basis of the hypothesized distances in difficulty between D, S, and CE problems. Rule I children should show a large distance between D on the one hand and S and CE on the other hand (D—S–CE), Rule II children should show a large distance between D and S on the one hand and CE on the other hand (D–S—CE), and finally, Rule III and Rule IV children should show a smaller distance between D and S on the one hand and CE on the other hand (D–S–CE).

Method

Participants. The data (generously shared and described more fully by van Maanen, Been, & Sijtsma, 1989) consisted of responses to Siegler-type balance beam problems by 484 students in Grade 7 or 8.

Indicators. The presentation of analyses will be restricted to a comparison between two kinds of problems: D and S. Five D problems and five S problems were considered. Results were similar for comparisons between the other pair of consecutive problems (S and CE) and among all three problems (D, S, and CE).

Manifest categories. Students who scored 0–5 were assigned to the first stage (Rule I level), and those who scored 6–10 were assigned to the second stage (Rule II level). This method of defining manifest categories is an example of segmentation. More sophisticated methods of defining categories (e.g., using latent saltus class probabilities) can also be applied (Wilson, 1989).

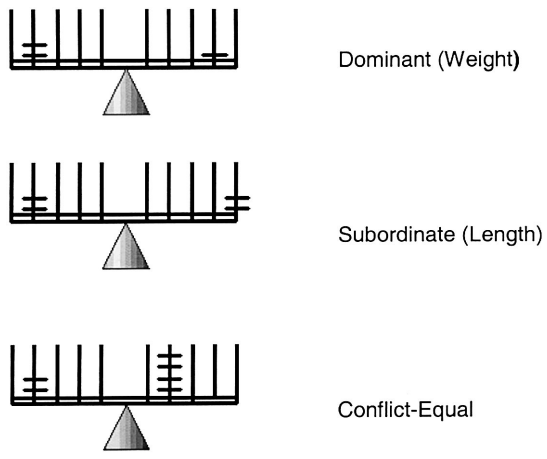

Figure 5. Siegler's (1981) balance problems for rule assessment.

Results

For the analyses we again estimated the Dimcat models, relying on the sequential strategy that was explained earlier. To begin, we estimated the QUAL1&2–HET model, the QUAL2–HET model, and the QUAN–HET model. In all three cases we were confronted with estimation problems: (a) extreme parameter estimates with extreme standard errors and (b) negative variances. All estimated models with either indicator-dependent discriminations, different variances for the two groups, or both, gave results that looked degenerated in one way or another. Therefore, we restricted the models to have equal indicator discriminations and a variance of one. The overall discrimination instead of the variance became a parameter. The crucial aspect to test was whether there were differences between the two groups with respect to the location of the problems, and if so, whether these differences could be explained with a saltus parameter for one type of tasks (in our case, increasing the distance between the S and the D problems).

Therefore, we tested three models: (a) a QUAL2–HET model with one task-independent overall degree of discrimination and with a person variance of 1 in both groups, (b) a QUAN–HET model with the same restrictions, and (c) a saltus model with a δ_S for the expected jump for the problems requiring that the subordinate dimension be used. The corresponding deviance values were 2,441.9, 2,729.9, and 2,448.5, respectively. The corresponding AIC and BIC values were 2,483.9 and 2,571.7 (QUAL2–HET), 2,753.9 and 2,804.1 (QUAN–HET), and 2,474.5 and 2,528.9 (saltus model), respectively. The likelihood ratio test comparing the restricted QUAL2–HET with the restricted QUAN–HET was statistically significant, $\chi^2(9) = 288.0$, $p < .001$, but when the saltus model was compared with the QUAL2–HET model, the difference in goodness of fit was not statistically significant, $\chi^2(8) = 6.6$, $p > .10$. The saltus model seemed to capture all qualitative differences between the two groups. It was also the best model with respect to the AIC and BIC. The estimates of δ_S indicated the size of the jump of the S items for the Rule II group. The estimated jump from the Rule I to the Rule II level was −4.856 (0.337), which was statistically significant given its standard error. The S problems were drastically easier at the Rule II level than at the Rule I level. No other differences were needed to approach the restricted QUAL2–HET model, so we concluded that the D tasks were equally easy for both groups.

After the assessment of between-category differences, we tested for within-category differences, in line with the vertical axis of Dimcat. The saltus model with homogeneity yielded a deviance of 2,507.5, with AIC and BIC values of 2,531.5 and 2,609.3, respectively, and a statistically significant (conservative) likelihood ratio test when compared with the corresponding model with heterogeneity, $\chi^2(1) = 59$, $p < .001$. The goodness of fit could largely be improved, however, when the discrimination for the Rule I level was fixed to zero (implying homogeneity in one group). The resulting deviance of 2,428.6 was also better than that of the corresponding full heterogeneity model. The heterogeneous model was in fact the best model of all those that could be estimated with good results. The AIC and BIC values were 2,454.6 and 2,509.0, respectively. Because the overall discrimination for the Rule II group was statistically significant, 1.551 (0.116), we concluded that there was homogeneity at the Rule I level and heterogeneity at the Rule II level. This finding was interesting, because it was the first time among our three applications that a manifest category turned out to be homogeneous.

We replicated the comparisons above for the S and CE problems, and also for the D, S, and CE problems (in the latter case, using a segmentation that yielded three manifest categories when all three kinds of problems were analyzed). The results for D and S problems replicated the above results, meaning that the difference was again qualitative and that again the manifest saltus model could explain this qualitative difference. For S and CE, one saltus parameter was again needed. To fit the data from the D, S, and CE problems, two saltus parameters were needed, one for the difference between D and S and one for the difference between D and CE.

Discussion

The findings show that development cannot be fully described by quantitative differences—there is a strong effect of student group (i.e., stage) on problem locations. This makes a Type 1 model with a saltus restriction the best model for the kind of development studied in this application. The latter result is not trivial, given that nothing in the formal way we performed the segmentation favored latent qualitative differences.

Another interesting finding is that the stages (or rule assessment classes) as defined by our segmentation rule are heterogeneous at the manifest level but not necessarily at the latent level. The Rule II stage seems to exhibit the microsequence phenomenon noted by K. W. Fischer et al. (1984), but the Rule I stage does not. A speculation to explain this result is that each stage shows the so-called microsequence phenomenon, implying within-stage quantitative development until a homogeneous end-state within the stage is reached, followed by a qualitative jump to the next stage, where again within-stage quantitative development occurs. The results can be explained by assuming that the Rule I students have reached the end-state of the Rule I level and that the other students are at different points of their quantitative development with respect to Rule II.

General Discussion

The first important result of the three applications is that all but one of the manifest categories that were defined on the basis of expert judgment or segmentation are heterogeneous, not just at the manifest level but also at the latent level. In principle, heterogeneity at the manifest level can originate from stochastic processes based on a homogeneous latent structure, with all persons in a manifest category being concentrated at one point in the latent structure, as shown in the simulation study for Application 2. This is the common case in the cognitive studies on categories and concepts: no latent continuum but only a manifest continuum (no internal category structure, no correlated features). In the present applications, by contrast, heterogeneity also occurred on the latent level, as indicated by differences in person locations. The only exception is the Rule I group in the developmental application. Thus, what some would consider categories on the basis of expert judgment or segmentation would seem to be rather heterogeneous entities. This result feeds back into the cognitive study of categories and concepts, and it is consistent with a need for giving more

attention to within-category structure, as expressed by Murphy (2002).

The second important result is that heterogeneity, even when captured by a descriptive dimension, does not necessarily imply that the manifest categories are only quantitatively different. To a small extent in Application 1 and to a large extent in Application 3, there was evidence for qualitative differences. Thinking of manifest categories as being dimension-like and reflecting qualitative differences may seem contradictory, but as we have shown qualitative differences and heterogeneity relate to different features of what it means to be dimension-like. In this situation, the use of the saltus parameters gives us a way to describe qualitative differences for a dimension-like structure.

The third important result is that when the differences are quantitative, the abruptness of the difference can be investigated at the latent level, so that one need not rely on the distribution of manifest variables, such as sum scores. In particular, in Application 2, in which quantitative differences were found, the manifest distribution and the latent distribution were both clearly bimodal, but this correspondence is not guaranteed, as shown by Grayson (1987). This was corroborated in our simulation study, showing that a structure with two categories with latent homogeneity can generate a smooth distribution, albeit one that can be identified as an artifact when the appropriate statistical tests are performed.

The fourth important result is that qualitative differences between manifest categories can sometimes be captured in a simple way. This is either because the qualitative differences are only minor (as in Application 1) or because a simple principle applies (as in Application 3). The latter is of special interest, because it allows one to test a theory of qualitative differences. In Application 3, the theory is Piaget's theory of cognitive development.

It is of interest to note that in our applications a large variety of latent structures were found, often with strong evidence against alternative structures. In all cases, we started from a rather simple manifest categorical variable, based on either expert judgment or segmentation. The implication of our findings is that manifest categories can differ a lot in their underlying structure. Without an investigation such as we conducted, one would perhaps not be aware of the quite different underlying status of the categorical variable one is using.

The differences between the different types of structure we found often turned out to be quite drastic: in all cases when within-category homogeneity versus heterogeneity was considered and also in Application 3 with respect to qualitative differences. Looked upon from this practical viewpoint, differentiating between the different types of structures was often not a problem. Although the issue of differentiating power remains an important one, it was shown in two simulation studies that for the kinds of differentiation that are relevant in our applications, the modeling approach we followed has good differentiation power and that modeling can correctly differentiate what the eye cannot.

Our approach hinges on the indicators that are selected, on the method of observation (e.g., ratings), and on the alternative manifest categories. For the study of personality disorders, the selection of the indicators was rather self-evident, given that both the indicators (symptoms) and the manifest categories (diagnoses) were based on the *DSM–IV*. For the study of attitudes, several alternatives were available. We could have referred to the circumstances of the crimes and to characteristics of the criminals, and one cannot tell whether these would have yielded the same results. For the study of cognitive development, the indicators certainly make sense, given that they are well-known tasks from this domain of study, but alternative tasks have been used. Perhaps the most severe limitation is that ratings were used in Applications 1 and 2, so that a cognitive bias may have affected the results. The conclusions must therefore be stated in terms of the manifest categories as used by raters. The situation is different for the developmental application, in which objective data were used. The choice of alternatives to a reference category is also an important issue. In some cases the choice is evident, as for the application on attitudes toward capital punishment and for the developmental application. However, for the personality disorder study, the category of people without any personality disorder would be a meaningful alternative category. The true nature of a category does not depend on the alternative categories it is compared with, but the alternative categories are an important methodological feature that restricts what one can or cannot find. For example, we believe that before one can come to a well-founded conclusion on personality disorders, it seems worthwhile to compare a given disorder with alternative disorders and with normality. One should also realize that our conclusions are restricted to pure personality disorders. Although we had reasons to use only pure categories, it prevents us from generalizing the results to the disorder categories as a whole.

Our approach is based on nonlinear mixed models for categorical data, and as such it is a very broad one, encompassing most IRT models and more. Analogous approaches can be developed rather easily for continuous data and for latent categories, but such developments would have a different scope. Instead we opted for bringing in another approach that is directed toward manifest categories and categorical indicators, one that is used in cognitive psychology: the generalized context model.

The link we made with the cognitive study of concepts and categories can be considered a mutually inspiring one. Our applications point to the need to include within-category heterogeneity and structure in studies on the cognitive representation of categories. In principle, one can analyze an Element $\times$ Feature matrix with elements from different categories, in the same way we did. On the other hand, the cognitive models are a good basis to investigate the way raters (experts and lay persons) come to a category-like decision on other persons or themselves. The cognitive models should be tested more for heterogeneous manifest categories, given that our results differ from those obtained with stimuli from categories without an internal structure (without correlated features).

We believe the approach we have formulated and applied is rather general and workable. It completes several other approaches, which can be deemed more specialized in one or another aspect of the concept of category-likeness. For example, the taxometric approach is specialized in detecting discreteness between latent categories along a dimension, and it concentrates on pairs of categories. Another example are methods to investigate factorial equivalence in its limited sense (checking only the factor loadings), which concentrate on discrimination equivalence, one aspect of qualitative versus quantitative differences. We do not claim that our framework is all-encompassing, but we believe that there is not just one feature that is distinctive for category-likeness, and that the metacategory of category-likeness is itself polythetic, as most categories are. It was our aim to leave freedom for such a poly-

thetic view of category-likeness, and that room was needed to explain our data.

References

Acton, G. S., Kunz, J. D., Wilson, M., & Hall, S. M. (in press). The construct of internalization: Conceptualization, measurement, and prediction of smoking treatment outcome. *Psychological Medicine.*

Acton, G. S., & Zodda, J. J. (in press). Classification of psychopathology: Goals and methods in an empirical approach. *Theory & Psychology.*

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

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

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43,* 567–573.

Beauchaine, T. P., & Beauchaine, R. J., III (2002). A comparison of maximum covariance and k-means cluster analysis in classifying cases into known taxon groups. *Psychological Methods, 7,* 245–261.

Beauchaine, T. P., & Waters, E. (2003). Pseudotaxonicity in MAMBAC and MAXCOV analyses of rating-scale data: Turning continua into classes by manipulating observer's expectations. *Psychological Methods, 8,* 3–15.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541–561.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison-Wesley.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Cantor, N., Smith, E. E., French, R. D., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology, 89,* 181–193.

Carson, R. C. (1991). Dilemmas in the pathway of the *DSM–IV. Journal of Abnormal Psychology, 100,* 302–307.

Clarkin, J. F., Widiger, T. A., Frances, A., Hurt, S. W., & Gilmore, M. (1983). Prototypic typology and the borderline personality disorder. *Journal of Abnormal Psychology, 92,* 263–275.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Coyne, J. C. (1994). Self-reported distress: Analog or ersatz depression? *Psychological Bulletin, 116,* 29–45.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43,* 171–192.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience, 1,* 77–94.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Orlando, FL: Harcourt Brace.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Endler, N. S., & Kocovski, N. L. (2002). Personality disorders at the crossroads. *Journal of Personality Disorders, 16,* 487–502.

Fischer, G. (1992). The 'saltus model' revisited. *Methodika, 6,* 87–98.

Fischer, K. W., Pipp, S. L., & Bullock, D. (1984). Detecting discontinuities in development: Methods and measurement. In R. N. Emde & R. Harmon (Eds.), *Continuities and discontinuities in development* (pp. 95–121). Norwood, NJ: Ablex.

Flett, G. L., Vredenburg, K., & Krames, L. (1997). The continuity of depression in clinical and nonclinical samples. *Psychological Bulletin, 121,* 395–416.

Frances, A., Widiger, T., & Fyer, M. R. (1990). The influence of classification methods on comorbidity. In J. D. Maser & C. R. Cloninger (Eds.), *Comorbidity of mood and anxiety disorders* (pp. 41–59). Washington, DC: American Psychiatric Press.

Gangestad, S. W., Bailey, J. M., & Martin, N. G. (2000). Taxometric analyses of sexual orientation and gender identity. *Journal of Personality and Social Psychology, 78,* 1109–1121.

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349.

Gangestad, S. W., & Snyder, M. (1991). Taxometric analysis redux: Some statistical considerations for testing a latent class model. *Journal of Personality and Social Psychology, 61,* 141–146.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.

Golden, R. R., & Meehl, P. E. (1979). Detection of the schizoid taxon with MMPI indicators. *Journal of Abnormal Psychology, 88,* 217–233.

Goodman, L. A. (1972). A general model for the analysis of surveys. *American Journal of Sociology, 77,* 1035–1086.

Grayson, D. A. (1987). Can categorical and dimensional views of psychiatric illness be distinguished? *British Journal of Psychiatry, 151,* 355–361.

Green, B. F. (1952). Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association, 47,* 71–76.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika, 55,* 477–494.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34,* 686–708.

Haslam, N. (1997). Evidence that male sexual orientation is a matter of degree. *Journal of Personality and Social Psychology, 73,* 862–870.

Haslam, N. (2002). Natural kinds, practical kinds, and psychiatric categories. *Psycoloquy, 13*(001). Retrieved January 4, 2002, from http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?13.1

Haslam, N., & Beck, A. T. (1994). Subtyping major depression: A taxometric analysis. *Journal of Abnormal Psychology, 103,* 686–692.

Haslam, N., & Cleland, C. (2002). Taxometric analysis of fuzzy categories: A Monte Carlo study. *Psychological Reports, 90,* 401–404.

Haslam, N., & Ernst, D. (2002). Essentialist beliefs about mental disorders. *Journal of Social and Clinical Psychology, 21,* 628–644.

Haslam, N., & Kim, H. C. (2002). Categories and continua: A review of taxometric research. *Genetic, Social, and General Psychology Monographs, 128,* 271–320.

Hidegkuti, I., & De Boeck, P. (2004). *The differentiation of Dimcat models: A simulation study.* Unpublished manuscript, K. U. Leuven, Belgium.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171–189.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Kass, F., Skodol, A. E., Charles, E., Spitzer, R., & Williams, J. B. W. (1985). Scaled ratings of *DSM–III* personality disorders. *American Journal of Psychiatry, 142,* 627–630.

Kelderman, H., & Steen, R. (1993). LOGIMO [Computer software]. Groningen, the Netherlands: ProGAMMA.

Kiers, H. A. L. (1990). *SCA: A program for simultaneous analysis of variables measured in two or more populations* [Computer software and manual]. Groningen, the Netherlands: ProGAMMA.

Kim, N. S., & Ahn, W. K. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General, 131,* 451–476.

Kofsky, E. (1966). A scalogram study of classificatory development. *Child Development, 37,* 191–204.

Komatsu, L. U. (1992). Recent views of conceptual structure. *Psychological Bulletin, 112,* 500–526.

Korfine, L., & Lenzenweger, M. F. (1995). The taxonicity of schizotypy: A replication. *Journal of Abnormal Psychology, 104,* 26–31.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* University of Chicago Press.

Lenzenweger, M. F. (1999). Deeper into the schizotypy taxon: On the robust nature of maximum covariance analysis. *Journal of Abnormal Psychology, 108,* 182–187.

Lenzenweger, M. F., & Korfine, L. (1992). Confirming the latent structure and base rate of schizotypy: A taxometric analysis. *Journal of Abnormal Psychology, 101,* 567–571.

Lilienfeld, S. O., & Marino, L. (1995). Mental disorder as a Roschian concept: A critique of Wakefield's harmful dysfunction analysis. *Journal of Abnormal Psychology, 104,* 411–420.

Livesley, W. J., Jackson, D. N., & Schroeder, M. L. (1992). Factorial structure of traits delineating personality disorders in clinical and general population samples. *Journal of Abnormal Psychology, 101,* 432–440.

Livesley, W. J., & Schroeder, M. L. (1990). Continua of personality disorder: The *DSM–III–R* Cluster A diagnoses. *Journal of Nervous and Mental Disease, 178,* 627–635.

Livesley, W. J., Schroeder, M. L., Jackson, D. N., & Jang, K. L. (1994). Categorical distinctions in the study of personality disorders: Implications for classification. *Journal of Abnormal Psychology, 103,* 6–17.

Maesschalck, C. (1998). *A psychometric modelling framework for testing categorical and/or continuous aspects of the borderline, histrionic, and antisocial personality disorders.* Unpublished doctoral dissertation, K. U. Leuven, Belgium.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning & Verbal Behavior, 23,* 250–269.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models.* New York: Wiley.

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park, NJ: Sage.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389–390.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469–1481.

Medin, D. L., & Coley, J. D. (1998). Concepts and categorization. In J. Hochberg & J. E. Cutting (Eds.), *Perception and cognition at century's end: Handbook of perception and cognition* (pp. 403–439). San Diego, CA: Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Meehl, P. E. (1973). MAXCOV–HITMAX: A taxonomic search method for loose genetic syndromes. In *Psychodiagnosis: Selected papers* (pp. 200–224). Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meehl, P. E. (1979). A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment, 43,* 563–581.

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

Meehl, P. E. (1999). Clarifications about the taxometric method. *Applied & Preventive Psychology, 8,* 165–174.

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology, 113,* 39–43.

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127–181). New York: Wiley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Miller, M. B. (1996). Limitations of Meehl's MAXCOV–HITMAX procedure. *American Psychologist, 51,* 554–556.

Millsap, R. E., & Everson, M. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J., & Bock, R. D. (1989). PC–BILOG 3: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71.

Murphy, G. L. (2002). *The big book of concepts.* Boston: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Nestadt, G., Romanoski, A. J., Brown, C. H., Chahal, R., Merchant, A., Folstein, M. F., et al. (1991). *DSM–III* compulsive personality disorder: An epidemiological survey. *Psychological Medicine, 21,* 461–471.

Nosofsky, R. M., & Palmeri, J. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300.

Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105,* 58–82.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition & categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous

item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75–92.

Rothschild, L., Cleland, C., Haslam, N., & Zimmerman, M. (2003). A taxometric study of borderline personality disorder. *Journal of Abnormal Psychology, 112,* 657–666.

Ruscio, A. M., Borkovec, T. D., & Ruscio, J. (2001). A taxometric investigation of the latent structure of worry. *Journal of Abnormal Psychology, 110,* 413–422.

Ruscio, A. M., & Ruscio, J. (2002). The latent structure of analogue depression: Should the Beck Depression Inventory be used to classify groups? *Psychological Assessment, 14,* 135–145.

Ruscio, J. (2000). Taxometric analysis with dichotomous indicators: The modified MAXCOV procedure and a case removal consistency test. *Psychological Reports, 87,* 929–939.

Ruscio, J., & Ruscio, A. M. (2000). Informing the continuity controversy: A taxometric analysis of depression. *Journal of Abnormal Psychology, 109,* 473–487.

Sanislow, C. A., Grilo, C. M., Morey, L. C., Bender, D. S., Skodol, A. E., Gunderson, J. G., et al. (2002). Confirmatory factor analysis of *DSM–IV* criteria for borderline personality disorders: Findings from the Collaborative Longitudinal Personality Disorder Study. *American Journal of Psychiatry, 159,* 284–290.

SAS Institute. (1999). *SAS online doc* (Version 8) [Software manual on CD-ROM]. Cary, NC: Author.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sher, K. J., & Trull, T. J. (1996). Methodological issues in psychopathology research. *Annual Review of Psychology, 47,* 371–400.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46,* 1–84.

Skilling, T. A., Quincey, V. L., & Craig, W. M. (2001). Evidence of a taxon underlying serious antisocial behavior in boys. *Criminal Justice and Behavior, 28,* 450–470.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smits, T., Storms, G., Rosseel, Y., & De Boeck, P. (2002). Fruits and vegetables categorized: An application of the generalized context model. *Psychonomic Bulletin & Review, 9,* 836–844.

Storms, G., & De Boeck, P. (1997). Formal models for intra-categorical structure that can be used for data analysis. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 439–459). London: UCL Press.

Storms, G., De Boeck, P., Hampton, J., & Van Mechelen, I. (1999). Predicting conjunction typicalities by component typicalities. *Psychonomic Bulletin & Review, 6,* 677–684.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory & Language, 42,* 51–73.

Strube, M. J. (1989). Evidence for the type in Type A behavior: A taxometric analysis. *Journal of Personality and Social Psychology, 56,* 972–987.

Sutcliffe, J. P. (1993). Concept, class and category in the tradition of Aristotle. In I. Van Mechelen, J. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 35–65). London: Academic Press.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology.* Cambridge, MA: Harvard University Press.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Taylor, J. R. (1995). *Linguistic categorization: Prototypes in linguistic theory* (2nd ed.). Oxford, England: Oxford University Press.

Thissen, D. (1997). MULTILOG [Computer software]. Mooresville, IN: Scientific Software.

Trull, T. J., Widiger, T. A., & Guthrie, P. (1990). Categorical versus dimensional status of borderline personality disorder. *Journal of Abnormal Psychology, 99,* 40–48.

Tyler, L. K., Moss, H. E., Dunant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75,* 195–231.

Tyrer, P., & Alexander, J. (1979). Classification of personality disorders. *British Journal of Psychiatry, 135,* 163–167.

van Maanen, L., Been, P., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–287). New York: Springer-Verlag.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1994). OPLM [Computer software and manual]. Arnhem, the Netherlands: CITO.

Vredenburg, K., Flett, G. L., & Krames, L. (1993). Analogue versus clinical depression: A critical reappraisal. *Psychological Bulletin, 113,* 327–344.

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

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods, 1,* 300–321.

Waller, N. G., & Ross, C. A. (1997). The prevalence and biometric structure of pathological dissociation in the general population: Taxometric and behavior genetic findings. *Journal of Abnormal Psychology, 106,* 499–510.

Widiger, T. A. (1992). Categorical versus dimensional classification: Implications from and for research. *Journal of Personality Disorders, 6,* 287–300.

Widiger, T. A., & Shea, T. (1991). Differentiation of Axis I and Axis II disorders. *Journal of Abnormal Psychology, 100,* 399–406.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.

Wilson, M. (1993). The "saltus model" misunderstood. *Methodika, 7,* 1–4.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford, England: Blackwell.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). ACER Conquest: Generalized item response modelling software [Computer software]. Melbourne, Australia: Australian Council for Educational Research.

Zimmerman, M., & Coryell, W. H. (1990). *DSM–III* personality disorder dimensions. *Journal of Nervous and Mental Disease, 178,* 686–692.

Appendix

Estimation and Reparameterization

As a basis for the analysis we start from a general formulation of the model for all manifest categories to be considered and with the first category ($k = 1$) as the reference category:

$$\eta_{pik} = (\alpha_{i1} + \alpha'_{ik})\theta_{pk} - (\beta_{i1} + \beta'_{ik}) + \gamma_k \quad \text{(A1)}$$

and

$$\theta_{pk} \sim N(0, \sigma_k^2)$$

α_{i1} is the discrimination of indicator i for the descriptive dimension of manifest category 1,
α'_{ik} is the deviation of α_{ik} from α_{i1} ($\alpha'_{ik} = \alpha_{ik} - \alpha_{i1}$), so that $\alpha'_{i1} = 0$,
β_{i1} is the location of indicator i for the descriptive dimension of manifest category 1,
β'_{ik} is the deviation of β_{ik} from β_{i1} ($\beta'_{ik} = \beta_{ik} - \beta_{i1}$), so that $\beta'_{i1} = 0$, and γ_k is the effect of manifest category k, $\gamma_1 = 0$.

If the within-group variance (σ_k^2) is free and possibly different depending on the manifest category, then for one item the discriminations should be restricted to be equal in all manifest categories. Alternatively, one can fix the variance in one category, for example $\sigma_k^2 = 1$. In a similar way the location of one item must be restricted to be equal in all manifest categories if the means of theta are zero and a γ_k is used.

Equation A1 is the basis for all the analyses below. We illustrate the SAS PROC NLMIXED approach for the case of two categories and six binary indicators: Category 1 is the reference category, and category 2 functions as the contrast category.

To check whether differences in reliability can account for a lack of equivalence, we also estimate the models with equal indicator parameters (alpha and beta) for all manifest categories and with a free within-category variance or an overall within-category discrimination. When such a model fits the data, one may conclude that equivalence is tenable but with possible differences in reliability.

We now explain how the estimation can be set up with SAS PROC NLMIXED for a full Type 1 model. Later (in Alternative Parameterizations) we discuss some parameterization issues.

The PROC NLMIXED Statements

Data Set and Options

This part of the statements indicates the data set and options for the estimation. The method we recommend for the integration is Gaussian quadrature (method=gauss): either nonadaptive (noad) or adaptive (noadscale). Nonapdative is much faster than adaptive and gives good results unless only a few items are used, but adaptive is better. We used the nonadaptive method with 20 quadrature points (qpoints=20). The optimization technique we used is Newton-Raphson with line search (technique=newrap). Also the maximum number of iterations (maxiter) and maximum number of function calls in the optimization (maxfu) are specified.

```
PROC NLMIXED
data=dimcat method=gauss noad technique=newrap qpoints=20 maxiter=5000 maxfu=5000;
```

Initial Values

This part of the statements gives initial values to all fixed effect parameters: the discrimination parameters in the reference category (a11–a16) and the deviations in discrimination in the contrast category (a21–a26); the location parameters in the reference category (b11–b16) and the deviations in location in the contrast category (b21–b26); and gam for the overall between-category difference.

```
PARMS
a11=1 a12=1 a13=1 a14=1 a15=1 a16=1
a21=1 a22=1 a23=1 a24=1 a25=1 a26=1
b11=0 b12=0 b13=0 b14=0 b15=0 b16=0
b21=0 b22=0 b23=0 b24=0 b25=0 b26=0
gam=0;
```

Conditional Formula Construction for the Probability

The formula is constructed with design variables of two kinds (to be included in the data set): The category is indicated with c1: c1=1 for category 1 and c1=0 for category 2, while c2=1 for category 2 and c2=0 for category 1 (this could be simplified using only c1, but the simpler way is cognitively more complex). The items are indicated with in1–in6, so that in1=1 for Item 1, and in1=0 otherwise, and this convention is followed for the Items 2–6. The formula is constructed in three steps: (a) Line 1–4, (b) Line 5, and (c) Line 6. Note that for identification reasons, the location deviation in Category 2 is fixed to zero for the first item. For an alternative, see Alternative Parameterizations in this Appendix.

```
ai1=a11*in1 + a12*in2 + a13*in3 + a14*in4 + a15*in5 + a16*in6;
ai2=a21*in1 + a22*in2 + a23*in3 + a24*in4 + a25*in5 + a26*in6;
bi1=b11*in1 + b12*in2 + b13*in3 + b14*in4 + b15*in5 + b16*in6;
bi2= + b22*in2 + b23*in3 + b24*in4 + b25*in5 + b26*in6;
ex = exp (ai1*c1*theta + (ai1+ai2)*c2*theta − bi1*c1 − (bi1 + bi2)*c2 + gam*c2);
p = ex / (1+ex);
```

Definition of the Stochastic Component

The distribution is a Bernoulli distribution, and it can be activated with "binary (p)"—see below.

```
MODEL response ~ binary (p);
```

The Definition of the Latent Variable Distribution

In the last part of the statements, the latent variable (or random effect) is defined: its distribution and distributional parameters and the kind of entities over which it varies (i.e., persons):

```
RANDOM theta~normal (0,1) subject=persons;
```

Note that one can estimate the variance in one or in both groups (var1 and var2). Unless a discrimination value (at minimum one) is fixed in each group, one of the variances needs to have a fixed value. In a model with discrimination equivalence, one variance needs to be fixed at a given value, or the overall discrimination needs to be fixed. When variance parameters are estimated, these parameters should be initialized: For example, var1=1 and var2=1, and the RANDOM statement is adapted as follows (for the case both variances are estimated):

```
RANDOM theta~normal (0,var1*c1 + var2*c2) subject=persons;
```

Run

Finally, the program is told to start running.

```
RUN;
```

(*Appendix continues*)

Alternative Parameterizations

Regarding the parameterization we presented earlier, two remarks are to be made. First, an equivalent parameterization is the following: $\eta_{pik} = (\alpha_{i1} + \alpha'_{ik})\,(\theta_{pk} - (\beta_{i1} + \beta'_{ik}) + \gamma_k)$. This is actually a better parameterization in case there are differences in discrimination. In the earlier parameterization, β'_{ik} is an equivocal parameter, because it reflects both location and discrimination differences, whereas in the parameterization with the alphas as multiplication factors for all parameters, the two are separated. If one wants to have a nonconfounded location difference parameter estimate (called *bdif12*) with the previously described approach in SAS, one can use the following statement:

```
ESTIMATE 'bdif22' b22/(a21 + a22);
```

This statement applies to Item 2 and can be repeated for all items by changing the first index from 2 into 3, 4, and so forth. Standard errors are also given.

Second, if one wants to have an estimate of the mean difference between the categories, effect coding (Cohen & Cohen, 1983, Chapter 5) for the items is to be preferred for the β' parameters instead of dummy coding as for the β parameters. This means that the item-specific predictor has a value of $+1$ for the item in question, -1 for a reference item, and 0 for all other items. This way of coding is also to be preferred for the identification of items with location differences.

Received October 31, 2002
Revision received March 11, 2004
Accepted March 23, 2004 ■