The influence of naive causal theories on lay concepts of mental illness

NANCY S. KIM AND WOO-KYOUNG AHN
Yale University

Two experiments, incorporating both real-life (Experiment 1) and artificial (Experiment 2) stimuli, demonstrated that lay concepts of mental disorders can be reliably predicted from subjects' naive causal theories about those disorders. Symptoms that are deeper causes (X, where X causes Y, which causes Z) are more important in lay concepts than intermediate causes (Y), which in turn are more important than terminal effects (Z). In addition, symptoms that cause or are caused by other symptoms are more important in lay concepts than symptoms not participating in any causal relationships. Implications of these results for current models of categorization and for research on lay theories of mental disorders are discussed, and future directions for research are suggested.

Placing others into categories of mental illness is a task that most laypersons face at various times during ordinary life. A person working in a company might remark to her colleagues that a scrupulous coworker must have some sort of obsessive–compulsive disorder. Likewise, a college student may worry that his roommate could be suffering from depression. The current investigation examines laypersons' conceptual representations and theories of mental disorders and the effect of these representations on judgments of category membership.

The task of categorizing others in the mental illness domain is not unusual for laypersons and can be extremely important. Indeed, Furnham (1995) recently proposed that laypersons' theories of the causes of mental disorders and their degree of resemblance to clinical theories may be predictive of how cooperative patients are when undergoing treatment for mental disorders. For instance, Narikiyo and Kameoka (1992) recently found that Japanese Americans were more likely than Caucasian Americans to attribute mental illness to social, external factors, which presumably led Japanese Americans to neglect the availability of mental health services and attempt to solve psychological problems on their own. In addition, investigating the mechanisms by which lay theories of mental disorders influence categorization and decision making may have important implications for cognitive researchers. Indeed, a recent approach to categorization, discussed in the following

AMERICAN JOURNAL OF PSYCHOLOGY
Spring 2002, Vol. 115, No. 1, pp. 33–65

section, posits that lay theories in general are critical for making categorization decisions.

Causal lay theories and disorder concepts

We addressed two major questions in this investigation. First, one might question whether laypersons have theories about mental illness. One traditional view is that people represent disorders as prototypes (Cantor, Smith, French, & Mezzich, 1980; Elstein, 1988). This prototype view has also been adapted for practical use in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV;* American Psychiatric Association [APA], 1994). According to this view, mental disorders are represented as lists of independent features that are not causally connected to each other. For example, clinicians using the prototype-like *DSM-IV* (APA, 1994) criteria would diagnose someone with obsessive–compulsive personality disorder if the patient had any combination of four symptoms out of a list of eight. In this case, it does not matter how these symptoms are related to each other or how they are caused (unless one is specified as a necessary diagnostic criterion).

Instead, Furnham and colleagues have demonstrated that laypersons seem to have intuitive theories about the causes of mental illness. Moreover, these lay theories are elaborate, consensual, and moderately accurate with respect to academic theories (Furnham & Bower, 1992; Furnham & Lowick, 1984). For example, Furnham and Hume-Wright (1992) have shown that laypersons generally agree that anorexia nervosa is caused by a number of factors, such as the stresses of coping with adolescence and the influence of images of thinness in the media.

What we are concerned with in the present study is whether laypersons also have theories of mental illness at a more detailed, internal level. That is, we propose that laypersons have theories about how symptoms of a disorder are causally connected to each other in the same way that they have theories about how features of an everyday category are causally connected (Lunt, 1991; Murphy & Medin, 1985).

For instance, consider three of the *DSM-IV* (APA, 1994) diagnostic criteria for anorexia nervosa: "refuses to maintain minimal body weight," "intense fear of gaining weight or becoming fat," and "disturbance in the way in which one's body weight or shape is experienced." It is easy to imagine a layperson thinking that people who have these anorexia nervosa symptoms "refuse to maintain minimal body weight" *because* they have an "intense fear of gaining weight or becoming fat" and also that they have an "intense fear of gaining weight or becoming fat" *because* they have a "disturbance in the way in which one's body weight or shape is experienced."

Indeed, the recent categorization literature has suggested that concepts are represented like theories or causal explanations. Characterizing this theory-based approach to categorization, Murphy and Medin (1985, p. 289) suggested that "concepts are coherent to the extent that they fit people's background knowledge or naive theories about the world." In other words, naive theories are the glue that holds the features of a concept together in a cohesive package. Murphy and Medin (1985), Keil (1989), and Carey (1985) have argued that models relying on similarity (e.g., Nosofsky, 1984; Posner & Keele, 1968; Tversky, 1977) cannot adequately explain human categorization. One of the main reasons for this is that the similarity-based models assume, as *DSM-IV* (APA, 1994) criteria imply, that features of concepts are independent of each other. However, in real life this is rarely the case. Our concept of birds, for instance, is not just a list of features (e.g., "have wings, can fly, build nests in trees"), but is rather a structure of causally related features (e.g., "birds can build nests in trees because they can fly, and birds can fly because they have wings"). Likewise, in answer to the first question we posed, we hypothesize that laypersons have naive theories about how at least some symptoms of each mental disorder are causally connected.

Our second research question follows from the first. How do laypersons' causal theories about mental disorders affect their categorization decisions? Researchers of lay causal theories and diagnoses of mental disorders have assumed that having causal theories about mental illness and diagnosing mental illness are two issues that can be adequately addressed separately (Westermeyer & Wintrob, 1979a, 1979b; but see Blatt & Levy, 1998, Wakefield, 1998, and the *General Discussion* for a different view of expert clinicians' reasoning). In other domains, Lunt and his colleagues have extensively documented the perceived causal structure of such constructs as loneliness (Lunt, 1991), poverty (Heaven, 1994), examination failure (Lunt, 1988), personal debt (Lunt & Livingstone, 1991a), and savings (Lunt & Livingstone, 1991b) without describing specific ways in which these causal structures would affect categorization. According to the theory-based approach to categorization, lay theories have direct ramifications for categorization (Carey, 1985; Keil, 1989; Murphy & Medin, 1985). For instance, although they look like fish, whales are categorized as mammals because of our naive biological theories about how fish breathe, reproduce, and so on, as compared with mammals. Our second question is how laypersons' causal theories of mental illness (if they have them) affect their categorization decisions.

Although there are many ways in which background theories might influence categorization or diagnosis, the current study focuses on testing the causal status hypothesis in the domain of mental illness (Ahn,

1998; Ahn, Kim, Lassaline, & Dennis, 2000). According to the causal status hypothesis, when features are causally related to each other, a cause feature influences categorization more than its effect feature. This causal status effect has been empirically supported in domains other than that of mental disorders, such as artifacts and natural kinds.

For example, in one experiment conducted by Ahn et al. (2000) involving novel natural-kind categories, participants read about three characteristic features of a target category (e.g., animals called "roobans" tend to eat fruits, have sticky feet, and build nests in trees). Participants in the control condition received no further information about the target category. In contrast, participants in the experimental condition were told that one feature tends to cause the second feature, which in turn tends to cause the third feature. In the rooban example, they were told that eating fruits tends to cause roobans to have sticky feet because the fruit sugars are secreted through pores on the undersides of their feet, and that having sticky feet tends to allow roobans to build nests in trees because they can climb up trees easily with their sticky feet. All participants were then presented with three exemplars, each of which had two features characteristic of the target category and one noncharacteristic feature (e.g., an animal that likes to eat worms, has feet that are sticky, and builds nests in trees). Participants were asked to rate how likely it was that this animal is a member of the target category (e.g., a rooban). For participants in the control condition, likelihood ratings remained constant regardless of which feature the exemplar animal was missing. However, in the experimental condition, the likelihood judgments varied as a function of the missing feature's causal status. Specifically, when an exemplar was missing the target category's fundamental cause in the causal chain, the mean likelihood of being a target category member was lower than when an object was missing its intermediate cause in the causal chain, which in turn was lower than when an object was missing its terminal effect.

In the present experiments, we investigated the effect of naive theories on mental illness concepts by applying the causal status hypothesis to that domain. Therefore, we expected symptoms that cause other symptoms to be more influential in mental illness categorization than symptoms caused by other symptoms. For example, in anorexia nervosa, if a layperson thinks that a "disturbance in the way in which one's body weight or shape is experienced" causes "intense fear of gaining weight or becoming fat," which in turn causes "refusal to maintain minimal body weight," then we would expect this layperson to give the most weight in categorization to the symptom "disturbance in the way in which one's body weight or shape is experienced" and the least weight to "refusal to maintain minimal body weight."

Overview of experiments

We report two experiments addressing the issue of how lay theories affect the classification of mental disorders. In both experiments, participants were presented with hypothetical patients to categorize, and their causal background theories of each disorder were either measured (Experiment 1) or manipulated (Experiment 2). Experiment 1 incorporated four mental disorders from the *DSM-IV* (APA, 1994) as stimuli. To better control participants' theories about mental disorders, Experiment 2 made use of artificial categories (created by combining *DSM-IV* symptoms from different disorders) and artificial theories of how these symptoms are causally connected to each other.

EXPERIMENT 1

In Experiment 1, mental disorders taken directly from the *DSM-IV* (APA, 1994) were used as stimuli. For purposes of comparison with the *DSM-IV,* we focused on symptoms that are considered diagnostic criteria symptoms by the *DSM-IV.* As explained earlier, the manual specifies that the presence of a certain subset of any combination of these symptoms (diagnostic criteria) is sufficient for a diagnosis of that disorder.

The task was divided into two parts: a causal centrality task in which participants drew causal relationships between symptoms and assigned causal strengths to each relationship and a conceptual centrality task in which participants rated the importance of symptoms in diagnosis. We hypothesized that each symptom is conceptually central to the extent that other symptoms are dependent on it (i.e., to the extent that it is causally central).

This hypothesis has been implemented in a computational model by Sloman, Love, and Ahn (1998). We used this model to test the causal status hypothesis in Experiment 1. The model is based on the following equation:

$$c_{i,t+1} = \Sigma_j d_{ij} c_{j,t}$$

where d_{ij} is a positive number that represents how strongly feature j depends on feature i, and $c_{i,t}$ is the conceptual centrality of feature i at time t.[1] The formula states that the centrality of feature i is determined at each time step by summing across the centrality of every other feature multiplied by that feature's degree of dependence on feature i. In essence, the model predicts that a feature is central to the extent that other features depend on it. As a simplified example of how the model predicts conceptual centrality from dependency centrality, say that feature X causes feature Y with a strength of 3, and Y causes feature Z also

with a strength of 3. If the initial conceptual centrality value is set to 1, after two iterations (of the matrix multiplication; see the *Method* section for details) the conceptual centrality of X is 16, the conceptual centrality of Y is 7, and the conceptual centrality of Z remains at 1. Note that this pattern of results mirrors the results of Ahn et al.'s (2000) rooban experiment: The more causally central a feature is, the more influence it has on categorization judgments. Thus, this model is one way to implement the predictions of the causal status hypothesis. Furthermore, the model is useful in deriving predictions of the causal status hypothesis for very complex patterns of causal relationships, which Sloman et al.'s (1998) experiments suggested people do have.

METHOD

Participants

We operationally defined "laypersons" as people who have never had any formal course training in abnormal psychology. Twenty Yale University undergraduate students participated in this experiment to fulfill partial requirements for an introductory psychology course, which is a prerequisite for all psychology courses at Yale University.

Materials

Four *DSM-IV* (APA, 1994) disorders that seemed likely to be familiar to undergraduate participants were selected. These were two Axis I clinical disorders (anorexia nervosa and major depressive disorder) and two Axis II personality disorders (narcissistic personality disorder and obsessive–compulsive personality disorder). For each of the four disorders, a list of criterial symptoms and characteristic symptoms was compiled.

The criterial symptoms were those that made up the *DSM-IV* (APA, 1994) diagnostic criteria for each disorder (for major depressive disorder, symptoms were taken from the list of criteria for a major depressive episode). Anorexia nervosa has four possible *DSM-IV* criteria, obsessive–compulsive personality disorder has eight possible criteria, and a major depressive episode and narcissistic personality disorder have nine possible criteria each. Thus, there were 30 criterial symptoms in all (see Table 1 for a complete list).

For each criterial symptom, a question measuring conceptual centrality was developed, following the format used by Sloman et al. (1998), Medin and Shoben (1988), and Barton and Komatsu (1989), among others. The question asked, "If a patient is in all ways like a typical person with X EXCEPT that he or she does NOT have the symptom of Y, does the patient have X?" where X is one of the four mental disorders and Y is a criterial symptom of that disorder. For example, one of the questions was the following: "If a patient is in all ways like a typical person with anorexia nervosa EXCEPT that he or she does NOT have the symptom of refusal to maintain minimum body weight, does the patient have anorexia nervosa?" The participant's answer was collected on a rat-

Table 1. Participants' conceptual centrality ratings for the 30 criterial symptoms of Experiment 1

Disorder and symptom	Conceptual centrality
Anorexia nervosa	
Refusal to maintain body weight at or above minimal levels	39.1 (30.9)
Fear of being fat even when underweight	44.0 (26.1)
Disturbed experience of body shape or denial of the problem	38.0 (24.1)
Absence of the period (in women) for 3+ menstrual cycles	17.3 (22.1)
Obsessive–compulsive personality disorder	
So preoccupied with details, rules, etc. that the goal is lost	33.2 (24.1)
Perfectionism that interferes with task completion	43.1 (26.0)
Excessively devoted to work and productivity	26.5 (18.2)
Scrupulous and inflexible about matters of morality, ethics	28.1 (17.3)
Unable to discard old junk with no sentimental value	13.9 (9.1)
Feels that their way is the only right way of doing things	31.6 (20.6)
Miserly toward self and others	24.8 (25.1)
Rigidity and stubbornness	33.8 (21.7)
Major depressive episode	
Depressed mood	44.2 (29.8)
Lack of pleasure in daily activities	33.1 (26.7)
Decrease or increase in weight	8.8 (8.8)
Sleep disturbances	11.2 (8.4)
Restlessness or unusual slowness	19.7 (18.8)
Fatigue or loss of energy	18.5 (12.8)
Feelings of worthlessness or excessive guilt	25.6 (15.4)
Indecisiveness or difficulty concentrating	17.1 (12.0)
Ideas or plans of suicide or suicide attempts	13.6 (12.4)
Narcissistic personality disorder	
Grandiose sense of self-importance	44.6 (24.6)
Preoccupied with fantasies of success and power	27.0 (20.2)
Believes only the special and unique can understand him or her	29.4 (24.4)
Requires excessive admiration	40.0 (26.8)
Feels entitled to special treatment	32.6 (20.6)
Exploits others for personal gain	30.2 (22.1)
Lacks empathy for the feelings and needs of others	34.2 (25.6)
Envious of others or believes that others are envious of him or her	30.4 (21.7)
Arrogant or haughty	42.3 (27.2)

Note. Actual ratings are subtracted from 100, so higher scores correspond to greater conceptual centrality in this table. Standard deviations are in parentheses.

ing scale of 0 to 100 (where 0 = *definitely no* and 100 = *definitely yes*). Because there was one question constructed for each criterial symptom, there were 30 conceptual centrality questions in all.

Additional characteristic symptoms that were not considered to be diagnostic criteria by the *DSM-IV* (APA, 1994) were taken from the description of each disorder. We extracted as many characteristic symptoms from the descriptions as possible that were not redundant with other criterial or characteristic symptoms for that disorder. Fourteen characteristic symptoms were used for anorexia nervosa, six for obsessive–compulsive personality disorder, nine for a major depressive episode, and five for narcissistic personality disorder. Thus, there were a total of 34 characteristic symptoms.

Procedure

Each participant received two tasks: a conceptual centrality task and a causal centrality task. The conceptual centrality task was performed on a Power PC Macintosh computer and was programmed using PsyScope 1.1 (Cohen, MacWhinney, Flatt, & Provost, 1993). This task attempted to measure the conceptual centrality of each of the criterial symptoms. To ensure that the participants had an idea of what each disorder was like before judging conceptual centrality, a list of both criterial and characteristic symptoms for each disorder was printed on four separate sheets and stapled into a booklet to be read before the computer task began. There were six randomized orders of the lists. Participants were first asked to read through the sheets listing the symptoms of each of the four disorders. They were told that these were characteristic symptoms as reported by trained clinicians and that this read-through was to give them an idea of what those disorders were.

Next, they were informed that the computer task would consist of a series of questions about these disorders. They were then presented with the 30 conceptual centrality questions described in the *Materials* section.[2] The questions were blocked by disorder and randomized within blocks, and the order of blocks was also randomized. This task was self-paced. Subjects responded by pressing keys on a keyboard and had the opportunity to change their answers before pressing the "Return" key to enter them. However, once participants pressed the "Return" key, they could not go back to change or refer to their previous answers.

In the causal centrality task, participants were asked to draw causal relationships between both criterial and characteristic symptoms. Both criterial and characteristic symptoms were included so that the participants' causal structures could be measured as comprehensively and as accurately as possible (e.g., if a characteristic symptom X were not included, a criterial symptom Z could erroneously be measured as causally peripheral when participants believed it to cause characteristic symptom X). Participants received a booklet consisting of an instruction sheet and four pages with the name of one mental disorder in the center of each page. Paper-clipped to each page with the disorder name was a sheet of 1/2-inch × 2 5/8-inch stickers labeled with the criterial and characteristic symptoms of that disorder. There were three different randomized orders of each sheet of labels. The order of disorders was randomized in each booklet.

During the causal centrality task, participants were instructed that their task was as follows. They were to read through the symptoms of the disorder and affix them to the sheet around the name of the disorder, grouping the symptoms that seemed to be related to make it easier to draw causal arrows between them in the next step of the task. Symptoms were not placed in a prearranged order to eliminate any possible suggestive influence their relative positions might have on participants' responses. Then, if participants believed that one symptom of the disorder caused another symptom of the disorder, they were asked to draw an arrow between the two symptoms. This method, which was also used in Sloman et al. (1998), differs from that used by Lunt and his colleagues (e.g., Lunt, 1991), in which participants were asked to judge causal strengths for every possible link between features that are presented in a matrix format. Our preference for the free arrangement method was based on the concern that the matrix format might encourage participants to think of pairs of features rather than the concept as a whole.

To clarify the manner in which arrows should be drawn, participants were also given an example: "If you believe that the symptom 'afraid of abandonment' causes the symptom 'eagerness to please others,' you should draw an arrow between the two symptoms, as illustrated below." An illustration was provided. Participants were also asked to draw as many arrows as they thought necessary to reflect all the causal relationships between symptoms on the page. They then assigned numbers to each arrow indicating the strength of that causal relationship, on a scale of 1–5 (where 1 means "X very weakly causes Y" and 5 means "X very strongly causes Y"). Participants were also instructed to feel free to write in any symptoms that they felt were missing and to cross out listed symptoms that seemed to be irrelevant to the disorder. Finally, participants were urged to take as much time as they needed to carefully complete the task.

The order in which participants received the two tasks (conceptual centrality task and causal centrality task) was counterbalanced across participants. All experiments were conducted in groups of one to four participants.

RESULTS AND DISCUSSION

Because the conceptual centrality questions were in the negated form of "If a patient is in all ways like a typical person with X EXCEPT that he or she does NOT have the symptom of Y, does the patient have X?" these scores were calculated by subtracting each from 100 for clarity of presentation in this analysis. Therefore, the higher the number is, the more conceptually central the symptom is to the disorder.

Table 1 presents the mean conceptual centrality ratings of the 30 criterial symptoms. A notable result is that conceptual centrality ratings of criterial symptoms were highly variable. For instance, in the category of anorexia nervosa, "fear of being fat even when underweight" was judged to be conceptually central (44.0), whereas "absence of the period (in women) for 3+ menstrual cycles" was judged to be conceptually peripheral (17.3). Such results, at least in laypersons, seem at variance

with the assumption implicit for many disorders in the *DSM-IV* (APA, 1994) that criterial symptoms are given equal weight in diagnosis.

The results from the causal centrality task are summarized in Figure 1. This figure presents averaged causal strengths among symptoms with-

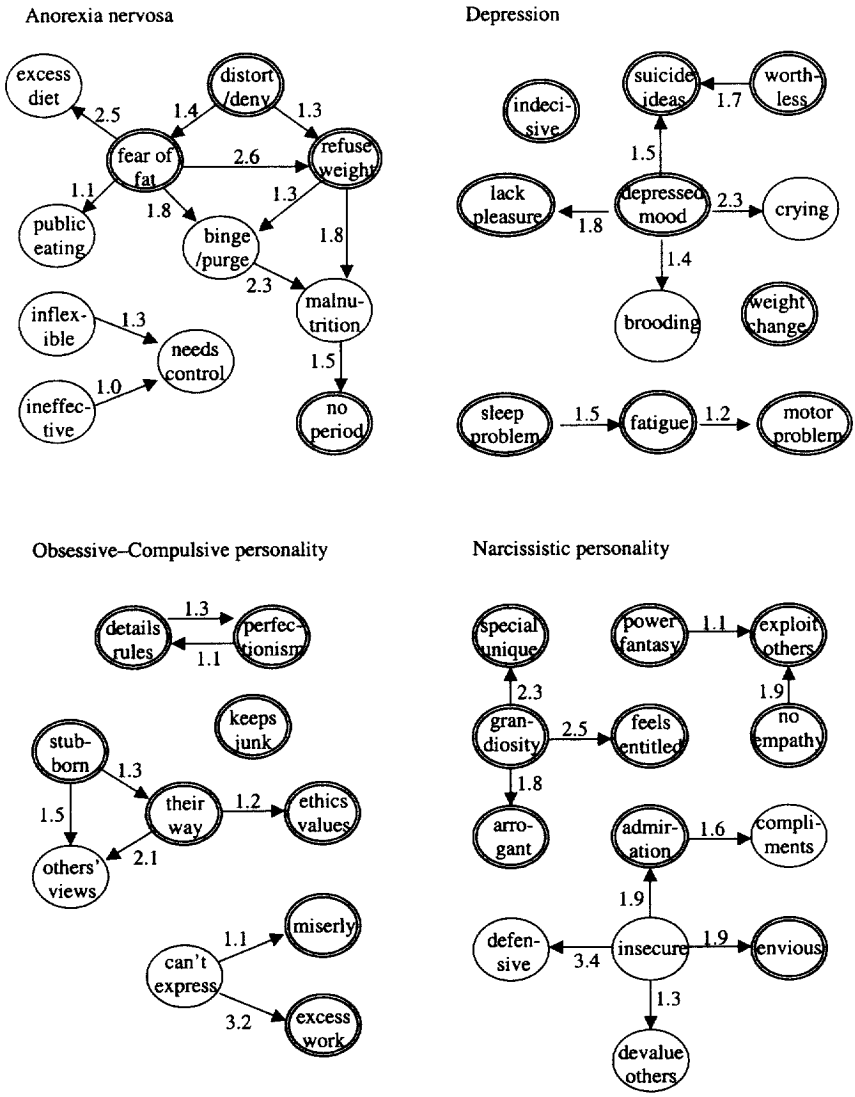

Figure 1. Averaged causal dependency structures in Experiment 1. Criterial symptoms are circled with a double line; causally unrelated characteristic symptoms are not shown.

in each disorder. These causal strengths were obtained by averaging participants' ratings of the strength of each causal link they drew.[3] The average causal strength scores could range from 0 to 5 because in the absence of a causal link between two features, 0 was used. For clarity of visual presentation, causal strengths lower than 1.0 were omitted from the figure (but were not omitted from the statistical analyses reported later in this article unless noted otherwise). The names of symptoms are abbreviated because of space limitations (full names of the criterial symptoms can be found in Table 1). The symptoms circled with double lines are criterial symptoms, and those circled with single lines are characteristic symptoms. One interesting result to notice is that even among criterial symptoms, causal centrality (as grossly indicated by the number of symptoms that a symptom causes and their strengths) seems highly variable. For instance, on average, participants believe that in anorexia nervosa, "fear of being fat even when underweight" causes many symptoms, including fear of eating in public, bingeing and purging, excessive dieting, and refusal to gain weight. However, "absence of the period (in women) for 3+ menstrual cycles," another criterial symptom for anorexia nervosa, was rarely judged to cause any other symptoms of that disorder. Figure 1 suggests that a conceptually central symptom (e.g., "fear of being fat even when underweight" in anorexia nervosa) is also causally central, and a conceptually peripheral symptom (e.g., "absence of the period (in women) for 3+ menstrual cycles") is also causally peripheral. This would be consistent with our primary hypothesis that a symptom is conceptually central to the extent that it causes other features. However, it is also apparent from Figure 1 that not all such pairwise comparisons, as in all of psychological research, were perfect one-to-one correspondences. Therefore, we tested for statistical significance as follows.

To quantify the causal centrality of a symptom, we used Sloman et al.'s (1998) model (i.e., the equation described in the introduction of Experiment 1) to derive predictions of conceptual centrality based on the ratings obtained in the causal centrality task. These predictions of conceptual centrality were then compared with the participants' actual conceptual centrality ratings obtained in the experiment. More specifically, a pairwise dependency matrix for each participant and disorder was first determined from their responses in the causal centrality task. That is, the strengths participants assigned to the causal arrows constituted the cells of the matrix. For each disorder, the matrices were averaged over all participants to yield a single matrix. Model-predicted conceptual centrality ratings were set to the initial arbitrary value of 0.5, following the procedure of Sloman et al. (1998).[4] The matrix multiplication was performed repetitively until the Spearman rank correlation

of the model-predicted conceptual centrality ratings and the conceptual centrality ratings given directly by the participants converged to its terminal stable value.

The main analyses were conducted using the procedure of Sloman et al. (1998). Rank correlations between the predicted and actual conceptual centrality ratings for each feature showed that the two factors were indeed positively correlated, $r_s = .73$.[5] Broken down by item, all correlations were positive, indicating that the effect did not result from any single disorder: $r_s(4) = .80$, $p = .2$ for anorexia nervosa; $r_s(9) = .73$, $p = .004$ for major depressive disorder; $r_s(8) = .86$, $p < .03$ for obsessive–compulsive personality disorder; $r_s(9) = .40$, $p < .3$ for narcissistic personality disorder (Figure 2).[6] These results are comparable to those of

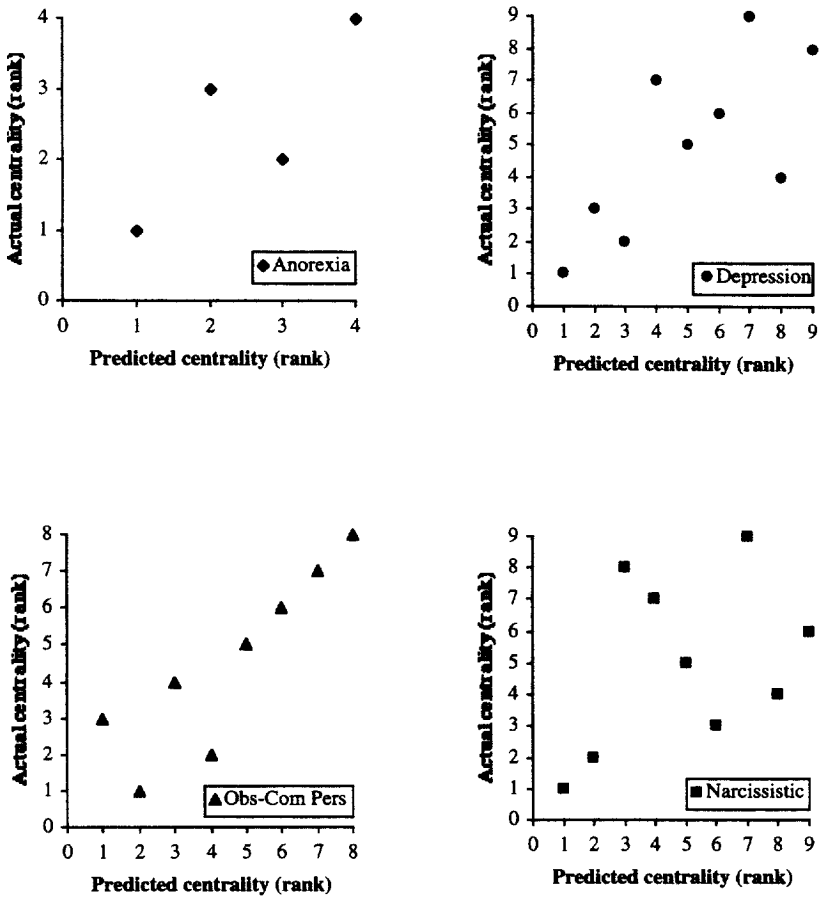

Figure 2. Rank-ordered correlations between model-predicted and actual given conceptual centrality in Experiment 1, broken down by disorder

Sloman et al. (1998, Study 2), who investigated other domains (i.e., guitars, apples, robins, and chairs), and support our hypothesis that features are conceptually central to the extent that they cause other features in the domain of mental illness.

It might be argued that this positive correlation simply demonstrates that participants were trying to be consistent across the two tasks, not that causal and conceptual centrality are related per se. That is, a participant might have given a high conceptual centrality rating on a certain symptom simply because he or she initially drew many strong causal relationships around this symptom. Likewise, a participant might have drawn many causal relationships around a certain symptom merely because he or she initially gave a high conceptual centrality rating on that symptom. To address this possibility, a between-subjects analysis was performed using only the data from whichever task participants performed first. The mean correlation across categories in this analysis was $r_s = .78$. Again, this clearly positive correlation did not result from any one disorder, $r_s(4) = .95$, $p = .05$ for anorexia nervosa; $r_s(9) = .92$, $p = .001$ for major depressive disorder; $r_s(8) = .42$, $p = .3$ for obsessive–compulsive personality disorder; $r_s(9) = .32$, $p < .4$ for narcissistic personality disorder. Compared with the within-subjects analysis described earlier, the clinical disorders (anorexia and depression) have higher correlations in these between-subjects analyses, whereas the personality disorders (obsessive–compulsive personality and narcissistic personality) have lower correlations. This pattern may be indicative of participants' having similar causal theories of each of these clinical disorders but very different theories of these personality disorders. We will return to this issue in the *General Discussion*.

The main analysis just described included all causal links marked by at least one participant (as was done in Sloman et al., 1998). We ran an additional analysis that considered causal links to be present only if they were reported with high consensus across participants. In this analysis, we included only the causal links that, collapsed over participants, were significantly greater than zero, $p < .05$, as determined by one-sample t tests. (One-tailed tests were used because causal ratings could not be lower than zero.) Rank correlations between the predicted and actual conceptual centrality ratings for each feature showed that the two factors were indeed positively correlated, $r_s = .68$. Broken down by item, all correlations were positive, again indicating that the effect did not result from any single disorder, $r_s(4) = .40$, $p = .6$ for anorexia nervosa; $r_s(9) = .73$, $p < .03$ for major depressive disorder; $r_s(8) = .81$, $p < .02$ for obsessive–compulsive personality disorder; $r_s(9) = .70$, $p < .04$ for narcissistic personality disorder.

Indeed, participants seem to have agreed among themselves as to which symptoms are more causally central than others. Kendall's co-

efficient of concordance (Marascuilo & McSweeney, 1977) was calculated from the rank orderings of centrality scores predicted by Sloman et al.'s model within each participant and within each disorder. In all four disorders, the correlation coefficients were significantly positive, W= 0.24, 0.23, 0.46, 0.50 for obsessive–compulsive personality disorder, anorexia, depression, and narcissistic personality disorder, respectively; all $ps < .001$.

Finally, to ensure that the results were not driven by the data of only a few participants, we also examined the results broken down by participant. The analyses were conducted in the same manner as the main analyses except that a separate analysis was run for each participant. Individually, 19 of 20 participants showed a positive correlation between predicted and actual conceptual centrality overall. Broken down by item, correlations were positive for a clear majority of participants in all disorders (15 out of 20 for anorexia nervosa, 16 out of 20 for major depressive disorder, 17 out of 20 for obsessive–compulsive personality disorder, 15 out of 20 for narcissistic personality disorder).

Summary

Experiment 1 presents a number of important findings concerning laypersons' conceptual representations of mental disorders. First, laypersons seem to have richly structured causal theories about mental disorders, as shown in Figure 1. Second, Experiment 1 presents the first demonstration of how naive theories of mental disorders can predict laypersons' categorization decisions. In accord with our hypotheses, symptoms that undergraduates rated as being causally central were also conceptually central. These data replicate the causal status effect (Ahn et al., 2000) in the realm of mental disorders.

EXPERIMENT 2

It could be argued that the correlational analyses in Experiment 1 do not necessarily reflect the impact of causal knowledge per se on conceptual centrality judgments. For instance, people might feel compelled to make more causal links to and from a symptom if it is conceptually central (rather than because causal centrality actually affects conceptual centrality). To test whether causal centrality does indeed influence conceptual centrality, Experiment 2 was designed using artificial mental disorders for which the participants' causal background knowledge, and only causal knowledge, was manipulated. In real-life diagnosis, it seems likely that all kinds of causal structures are used. Similarly, laypeople in Experiment 1 constructed a number of different kinds of causal struc-

tures (see Figure 1). These included simple causal links (i.e., X → Y), longer causal chains (i.e., X → Y → Z), common cause structures (i.e., X → both Y and Z), and common effect structures (i.e., X and Y both → Z). For the sake of simplicity, in this experiment we focused only on three-step causal chains (X → Y → Z). (See Ahn & Kim, 2000, for a discussion of the causal status hypothesis in common effect structures.)

Experiment 2 also examined the effect of symptoms that are not involved in any causal relationships with other symptoms. We were interested in what type of influence, if any, these causally unrelated symptoms would have in diagnosis compared with symptoms involved in causal relationships. For instance, consider four symptoms of a mental disorder, X, Y, Z, and W, where the first three symptoms form a causal chain (X → Y → Z) and the last symptom, W, is causally unrelated. Would W affect diagnosis as much as the most fundamental cause (X), the intermediate cause (Y), or the most terminal effect (Z)? Alternatively, could it be that W affects diagnosis even less than Z? Previous studies testing the effect of causal status of features on categorization have not examined this issue, although it is important to do so because real-life concepts, including concepts of mental illness, sometimes have causally unrelated features. For instance, most people would find it difficult to explain why tires are black and how this feature is causally related to other features of a tire.

Some clues to answering this question of causally unrelated features come from the induction literature. According to Gentner's (1983, 1989) structure-mapping theory, relational features (statements taking two or more arguments; e.g., "x is smaller than y") are more important than attributes (statements taking only one argument; e.g., "x is yellow") in analogical inference. For instance, in the analogy "An atom is like the solar system," attributes such as "yellow," "hot," and "massive" for the sun are not useful in making the analogy and are therefore discarded. However, relational features such as "more massive than" and "revolves around" can be used to draw the analogy that electrons revolve around the nucleus in an atom as planets revolve around the sun in a solar system.[7] Lassaline (1996) provided evidence for Gentner's theory in category-based induction. In a similar vein, Billman and her colleagues showed that it is easier to learn a rule that links two features when the link is part of a system of correlations than when it occurs in isolation (e.g., Billman, 1989; Billman & Knutson, 1996). Given these findings, we hypothesized that features related to other features would be given greater weight in categorization than unrelated features. In the aforementioned example containing the causal chain (X → Y → Z) and a causally unrelated feature (W), we would expect each of symptoms X, Y, and Z to be given more weight in categorization than symptom W.

Experiment 2 used artificial stimuli to make a direct comparison of relational and causally unrelated symptoms.

In Experiment 2, participants were presented with six artificial mental disorders in turn. For each disorder, they read about its six symptoms and some background knowledge about the disorder and then were asked to answer a conceptual centrality question of the form used in Experiment 1 for each symptom. One of two possible types of background knowledge was given for each disorder. One type stated that three of the symptoms were causally connected (i.e., $X \rightarrow Y \rightarrow Z$) and that the other three symptoms are not causally connected to one another or to any other symptom. The second type of background knowledge served as the control, in which none of the six symptoms were causally connected to any other symptom. The noncausal background information was constructed so that it would rule out the possibility that participants might give greater weight to symptoms X, Y, and Z as opposed to the noncausal symptom W merely because X, Y, and Z were part of a group.[8] Specifically, in this information, the symptoms were given a common grouping factor, but the amount of causal information that participants could possibly extract from the information was minimized by the nature of its description (see the *Materials* section).

In summary, Experiment 2 tested two main hypotheses. First, it was predicted that a deeper cause (e.g., symptom X, when X causes Y, which causes Z) will be weighted more heavily in categorization of mental disorders than an intermediate cause (e.g., symptom Y), which in turn will be weighted more heavily than the terminal effect (e.g., symptom Z). Second, it was hypothesized that symptoms that are part of a causal chain will be weighted more heavily in categorization of mental disorders than symptoms thought to be causally unrelated.

METHOD

Participants

Forty Yale University undergraduate students participated in this experiment and other unrelated experiments for a payment of $7. Each participant was randomly assigned to one of the experimental conditions.

Materials and design

The stimulus materials were six artificial mental disorders, each made up of six symptoms taken from various disorders listed in the *DSM-IV* (APA, 1994). Symptoms were chosen randomly, with the exception that no two symptoms within an artificial disorder were taken from the same *DSM-IV* (APA, 1994) disorder. This measure was taken to minimize the amount of a priori information participants had about how the symptoms might fit together.[9] For exam-

ple, the six symptoms of the artificial disorder "Methinismus" were intense fear of gaining weight, recurrent unjustified suspicions of the spouse's infidelity, fake suicide attempts, perfectionism that interferes with task completion, unreasonable scorn for authority, and exaggerated startle response.

In the scenarios participants received, symptoms in each disorder were said to be related in either a causal or noncausal manner. When symptoms were causally related (e.g., intense fear of gaining weight causes recurrent unjustified suspicions of the spouse's infidelity, which in turn causes making fake suicide attempts), a plausible explanation was also given (e.g., "An intense fear of gaining weight causes these patients to have recurrent unjustified suspicions of their spouses' infidelity, because they fear having become fat and unattractive to their spouses. These recurrent unjustified suspicions, in turn, cause these patients to make fake suicide attempts aimed at making the spouse feel guilty about the supposed affair."). These plausible explanations were given because the symptoms of each disorder were chosen deliberately so that causal links would not be immediately obvious to the participants in the control condition. Therefore, simply saying that the causal links existed without explaining how would make the causal relationships insufficiently plausible. Implausible causal relationships would be a problem because we have shown elsewhere that the causal status effect does not and should not be expected to occur when participants do not believe the causal relationships given (Ahn et al., 2000).

When symptoms were not causally related, we provided background information that would allow participants to group the noncausal symptoms together without providing any information to them concerning causal relationships between the symptoms. These grouping factors were added to equate, as much as possible, materials between causal and noncausal relationships with respect to the number of relationships in which a symptom participates and the length of the background knowledge. An example is the following: "At a recent conference on mental illness, a speaker talking about intense fears of gaining weight was praised because she remained collected when the microphone broke. At a recent conference on mental illness, a speaker talking about recurrent unjustified suspicions of the spouse's infidelity was praised because she had charisma. At a recent conference on mental illness, a speaker talking about making fake suicide attempts was praised because she handled the audience's questions well." Six different grouping factors were generated (e.g., articles on each of the symptoms were published in a particular journal; patients with the symptoms were more likely than the general population to have particular digits in their Social Security numbers).

Using these causal and noncausal relationships, scenarios for three conditions were constructed: the first-cause, the last-cause, and the no-cause conditions. In the scenarios used for the first-cause condition, the first three symptoms were causally related and the last three symptoms were not causally related. In the scenarios used for the last-cause condition, the first three symptoms were not causally related and the last three causes were causally related. (See Appendix B for a sample scenario used for the last-cause condition.) We included both the first-cause and the last-cause conditions to ensure that any

effects found are caused not by the content of the features selected to be causal but rather by the causal status of the features per se. For instance, "perfectionism that interferes with task completion" is the deepest cause in the last-cause condition but a causally unrelated symptom in the first-cause condition. Therefore, the difference between the two conditions would demonstrate the effect of the causal status of symptoms regardless of symptom content.

Finally, in the no-cause condition, all six symptoms were causally unrelated. This condition allowed us to compare the impact of a causally unrelated symptom in the context of all causally unrelated symptoms (i.e., no-cause information) with the impact of the same causally unrelated symptom in the context of causally related symptoms (i.e., first- and last-cause information). That is, this condition enabled us to explore the possibility that the same causally unrelated symptom might have a different impact on diagnosis when there are other causally related symptoms in the same category than when there are not. (Appendix A reports a study conducted as a manipulation check to see whether the noncausal relationships with common grouping factors are indeed thought to be noncausal and whether the passages we intended as causally linked were indeed thought to be strongly causally linked by participants from the same population as the participants in the main experiment.)

In all three conditions, each disorder was presented in the following manner on a single page so that participants could consider all the information given while answering the questions (for a specific example, see Appendix B). The symptoms were listed under the header "Diagnostic criteria for Axis I disorder: [Name of the disorder]" and were also prefaced by a clause stating, "Three (or more) of the following symptoms have been present for at least 1 month." In the first-cause and last-cause conditions, arrows pointing from cause to effect were drawn between the symptoms said to cause one another in the background information. Below the symptoms, it was stated, "In the diagram above, an arrow between two symptoms indicates the direction in which one causes the other. No arrow between any two symptoms means that there is no causal relationship between them." Next, a sentence stating, "The following facts are true of the symptoms of [Name of the disorder]," was followed by one of the background condition paragraphs.

Finally, listed at the bottom of the page were conceptual centrality questions of the type used in Experiment 1. For each of the six symptoms in each disorder, participants were asked a question in the form of "If a patient is in all ways like a typical person with X EXCEPT that he or she does NOT have the symptom of Y, does the patient have X?" where X is one of the six mental disorders and Y is a symptom of that disorder. Participants gave their answers to each question on a scale of 0 to 100 (where 0 = *definitely no* and 100 = *definitely yes*). (See Appendix B for examples.)

Each participant received two disorders in the first-cause condition, two in the last-cause condition, and two in the no-cause condition. A Latin-square design was used to determine the combination of disorder (six different artificial disorders), background knowledge condition (first-cause, last-cause, or no-cause), and type of grouping factor (six different noncausal stories; recall that each background knowledge condition incorporates a noncausal grouping

factor for at least three symptoms of the disorder). Approximately equal numbers of participants were randomly assigned to each of the Latin-square conditions. In addition, the presentation order of the conceptual centrality questions (randomized for each disorder and each condition, and the reverse of the randomized orders) was counterbalanced.

Procedure

Participants were given a booklet containing the six disorders as described in the *Materials* section. Instructions for the experiment were printed on the cover page. Participants were told in the instructions that they would read a brief description of a mental disorder on each of the following pages. The instructions then stated that participants would be asked to imagine a patient who has all of these symptoms. This was to ensure that participants, having just learned the disorder, had some representation of what a perfect-example patient with the disorder was like. Participants were told that at the bottom of the page there would be six questions about hypothetical patients, that the participants would be asked to answer on a scale provided. Participants were told that there would be six different mental disorders in all and that they were free to ask the experimenter any questions. Participants then read each page and completed the conceptual centrality judgments for the hypothetical patients. The order of the six disorders was randomized across the participants.

To summarize the design and the procedure, each participant learned about six artificial mental disorders with six symptoms each. Each disorder was accompanied by a paragraph of background information about the relationships between the symptoms. Using abstract notations of $\rightarrow$ to indicate a causal relationship and "/" to indicate a noncausal relationship and numbers 1 to 6 to indicate the six symptoms, symptoms in the first-cause condition were $1 \rightarrow 2 \rightarrow 3 / 4 / 5 / 6$, symptoms in the last-cause condition were $1 / 2 / 3 / 4 \rightarrow 5 \rightarrow 6$, and symptoms in the no-cause condition were $1 / 2 / 3 / 4 / 5 / 6$. In each disorder, participants gave a rating of the conceptual centrality of each symptom.

RESULTS AND DISCUSSION

In this experiment we were concerned with whether causal background knowledge (compared with noncausal knowledge) increases the weight of a feature in making conceptual centrality judgments, such that a feature thought to be a cause is given more weight than a feature thought to be its effect. We were also interested to see whether the terminal effect feature in a causal chain would be given more weight in conceptual centrality judgments than a noncausal feature in the same category. As in Experiment 1, ratings from the conceptual centrality questions were inverted for ease of understanding so that in the analysis, the higher the number is, the more conceptually central the symptom is to the disorder.

The mean ratings for Experiment 2 are listed in Table 2. Overall, our hypothesis was supported. In the no-cause condition, the six symptoms did not differ. However, when the same symptoms were causally related, causal depth determined the membership likelihood (i.e., in the first-cause condition with 1 → 2 → 3 / 4 / 5 / 6, the conceptual centrality of symptom 1 > symptom 2 > symptom 3; in the last-cause condition with 1 / 2 / 3 / 4 → 5 → 6, the conceptual centrality of symptom 4 > symptom 5 > symptom 6). Furthermore, causally unrelated symptoms were significantly less conceptually central than even the most terminal effect in a causal chain (i.e., in the first-cause condition, symptom 6 < symptom 3; in the last-cause condition, symptom 3 < symptom 6). The following statistical analyses support this conclusion.

A 3 (condition: first-cause, last-cause, no-cause) × 6 (symptom: 1, 2, 3, 4, 5, 6) repeated-measures ANOVA revealed significant main effects of condition, $F(2, 78) = 4.27$, $p < .02$, and of symptom, $F(5, 195) = 6.85$, $p < .001$. These main effects should be interpreted in light of the significant interaction effect, $F(10, 390) = 27.92$, $p < .001$, because they resulted primarily from higher ratings of causally related symptoms in the first-cause and the last-cause condition. A one-way ANOVA was run in each condition to determine which conditions drove these effects. In the two causal conditions, symptoms were differently weighted: first-cause: $F(5, 195) = 27.80$, $p < .001$; last-cause: $F(5, 195) = 23.82$, $p < .001$. In contrast, symptoms were not differently weighted in the no-cause condition, $F(5, 195) = 1.07$, $p = .4$.

Planned comparisons were run in the two causal conditions to determine which pairs of symptoms differed in weight. In the first-cause condition, the deepest cause in the chain was rated as significantly more conceptually central than the intermediate cause (mean ratings of 55.0 and 48.2, respectively; $t(39) = 2.29$, $p < .03$). Similarly, in the last-cause

Table 2. Mean conceptual centrality ratings in Experiment 2

	Symptom					
Condition	1	2	3	4	5	6
First-cause	55.0 (28.0)	48.2 (26.3)	39.5 (28.0)	24.5 (18.7)	23.3 (18.0)	24.0 (19.3)
Last-cause	25.9 (21.2)	26.5 (18.5)	26.4 (18.8)	53.0 (25.0)	44.8 (24.1)	36.0 (23.8)
No-cause	30.3 (23.0)	28.5 (21.2)	28.9 (21.4)	28.9 (22.1)	29.8 (23.1)	31.6 (24.8)

Note. Actual ratings are subtracted from 100, so higher scores correspond to greater conceptual centrality in this table. Standard deviations are in parentheses. In the first-cause condition, symptom 1 is the deepest cause, symptom 2 is the intermediate cause, symptom 3 is the terminal effect, and symptoms 4, 5, and 6 are noncausally grouped. In the last-cause condition, symptoms 1, 2, and 3 are noncausally grouped, symptom 4 is the deepest cause, symptom 5 is the intermediate cause, and symptom 6 is the terminal effect. In the no-cause condition, all symptoms are noncausally grouped.

condition, the deepest cause was significantly more conceptually central than the intermediate cause (mean ratings of 53.0 and 44.8, respectively; $t(39) = 3.61$, $p = .001$). In both the first-cause and last-cause conditions, the intermediate cause was also rated as more conceptually central than the terminal effect of the chain (first-cause, mean ratings of 48.2 and 39.5, respectively, $t(39) = 2.99$, $p = .005$; last-cause, mean ratings of 44.8 and 36.0, respectively, $t(39) = 4.07$, $p < .001$). Finally, the terminal effect was rated as more conceptually central than the noncausal symptom in both conditions (first-cause, mean ratings of 39.5 and 24.5, respectively, $t(39) = 3.61$, $p = .001$; last-cause, mean ratings of 36.0 and 25.9, respectively, $t(39) = 2.49$, $p < .02$). An item analysis revealed that for each of the six disorders, these data patterns appeared consistently.

In addition, comparisons were made between the weight given to noncausal symptoms in the causal conditions, $M = 24.7$, and the no-cause condition, $M = 33.5$. More weight was given to noncausal symptoms in the no-cause condition than in the causal conditions, $t(23) = 2.42$, $p < .03$. That is, the same noncausal symptoms are perceived to be less conceptually central when there are other, more causally central features in the causal conditions.

Overall, Experiment 2 demonstrated that participants given causal information gave greater conceptual centrality to cause symptoms than to their effects. Effect symptoms were also given greater conceptual centrality than causally unrelated symptoms in the pooled causal information condition. In the noncausal information condition, in contrast, participants gave the same symptoms equal conceptual centrality.

One might be concerned with the use of the common grouping factors in the no-cause condition (e.g., speakers talking about symptoms 1, 2, and 3 were all praised by the audience in a conference). Although these grouping factors were added to maximally equate the no-cause and the two cause conditions except for the presence of causal relationships, the use of noncausal grouping factors might look artificial because they may seem to incorporate unlikely coincidences. Furthermore, introducing factors that seem irrelevant to diagnoses might have introduced an unexpected confounding variable. However, in a similar experiment reported in Kim (1999), no grouping factors were used in the no-cause condition, and participants were simply told that each symptom does not and is not caused by any other symptoms of the same disorder. Even with this alternative control, the same pattern of results was obtained.

GENERAL DISCUSSION

We found evidence in two experiments that laypersons' theories about how symptoms of mental disorders are causally related have a signifi-

cant impact on the classification decisions they make. These demonstrations were made using pre-existing *DSM-IV* (APA, 1994) mental disorders (Experiment 1) and experimentally developed mental disorders (Experiment 2).

More specifically, Experiment 1 showed that laypersons have rich naive causal theories about mental disorders and that each symptom affects categorization decisions to the extent that it causes other symptoms in the same category. Experiment 2 found that in an X → Y → Z causal chain, deeper cause symptoms (X) were more conceptually central than intermediate causes (Y), which were more conceptually central than terminal effects (Z), suggesting that causal centrality can affect conceptual centrality. Furthermore, Experiment 2 demonstrated that symptoms part of an X → Y → Z causal chain were rated as more conceptually central than a noncausal, noncaused W symptom. Because novel mental disorders were used to manipulate the causal relationships between symptoms, Experiment 2 established that lay theories about relationships between symptoms in a mental disorder can *determine* the importance of the symptoms when a layperson thinks about that mental disorder.

In the following sections we will discuss implications of these results for both categorization theories and clinical research and suggest some follow-up studies and future directions for research.

Implications for categorization theories

This investigation replicated the causal status effect reported by Ahn et al. (2000) with stimuli from a different domain and also gave general support for the model of Sloman et al. (1998) in the domain of mental illness. In addition, the current study also revealed several other novel findings pertinent to the categorization literature.

First, to our knowledge it is the first empirical demonstration that causally unrelated features (i.e., W symptoms) are given less weight in category membership decisions than features thought to be involved in causal relationships (including terminal effect features). Thus, the result provides another way of using causal background knowledge to constrain feature weights in categorization in addition to the one proposed by the causal status hypothesis.

This result also suggests at least one way of fine-tuning the Sloman et al. (1998) model. This model currently predicts that causally unrelated features and terminal effect features will have the same weight in categorization. For example, suppose that feature X causes feature Y with a strength of 3, Y causes feature Z with a strength of 3, and feature W is a causally unrelated feature in the same category as X, Y, and Z. If the initial conceptual centrality value is set to 1, after two iterations (of the matrix multiplication) the model predicts that the conceptual cen-

tralities of X, Y, Z, and W are 16, 7, 1, and 1, respectively. In the current study, however, participants gave different weights to terminal effects (Z) and causally unrelated features (W). We speculate that, after making modifications to the model, the data from Experiment 1 may be an even better fit. Future work will determine whether this is the case.

It should be further noted that this finding—that W symptoms, those that are not causally connected to any other symptoms, received less weight than symptoms that are causally related to other symptoms (i.e., X, Y, and Z)—was obtained even though these W symptoms were related in a noncausal way to other symptoms (i.e., U and V). Critically, these W symptoms were related to other symptoms in a noncausal manner (or what we termed as common grouping factors). Therefore, this finding suggests that not all relationships are treated equally, as has sometimes been implied in existing theories of induction (e.g., Gentner, 1989, but see Holyoak, 1985). Rather, these results imply that causal relationships may be more important for categorization than other grouping factors. However, this possibility must be more systematically tested in future studies. It is possible that this importance of causal relationships might be specific to the mental illness domain or that the noncausal grouping scenarios we used in Experiment 2 were not regarded seriously by the subjects.

The second novel finding of the current study in the context of categorization literature is that in Experiment 2, features not part of a causal group are given more weight when the other features in the category are also not part of any causal group than when the other features in the category are causally related to each other. That is, noncausal W symptoms were rated as more conceptually central when X, Y, and Z were not causally related than when $X \rightarrow Y \rightarrow Z$. This context effect cannot be accounted for by the current version of Sloman et al.'s model, which predicts that the conceptual centrality of W should remain as the same initial value in both cases. What this context effect suggests is that as one gains more knowledge about causal relationships between features, the features that do not have causal relationships to other features will have even less impact on categorization than they did before. For instance, a student just learning about what anorexia nervosa is might give little weight in categorization to the symptom "absence of the period for more than 3 months." However, the student might then give even less weight in categorization to that symptom as he or she develops his or her own theories about how the other symptoms of anorexia nervosa are causally related (provided that "absence of the period" is still not thought to be causally related to any other symptoms, as our data showed in Figure 1). It remains to be seen whether this context effect occurs in domains other than mental disorders.

There remains the possibility that response artifacts are an alternative explanation for this context effect. For example, in the causally related condition, judgments of noncausal factors could have been reduced as a result of a contrast effect, tending to push them away from and below judgments of the causally related factors. Or, in the noncausal condition, there might have been implicit demands to not give low ratings to all symptoms, subsequently pushing ratings of those symptoms up a little (that is, if the effect of the demand was distributed evenly across symptoms for the sample as a whole). In future research to directly test these hypotheses, one might want to use scales with more explicit phrases for each point on a scale (i.e., percentages) and also to explicitly encourage participants to make the ratings they feel are more appropriate, regardless of what they think the experimenter might be expecting.

Implications for the study of lay theories

In this article, we have been concerned with whether laypersons have theories about how symptoms of a disorder are causally connected to each other, and with the relationship between laypersons' causal theories and lay diagnostic criteria for classification into a mental illness category. The research we have presented here provides evidence that lay classification decisions concerning mental disorders are affected strongly by people's naive causal theories about those disorders.

In generalizing the current results to laypersons' real-life diagnosis of mental disorders, one might be concerned that the format of the dependent measures used in the current studies does not correspond to how laypersons classify people. In both studies, we measured the conceptual centrality of symptoms by asking participants to rate the likelihood that a person without a certain symptom would have a target category. That is, we measured the conceptual centrality of symptoms by negating them. Kim (1999) used dependent measures that are more similar to real-life diagnostic reasoning. Specifically, she gave participants descriptions of hypothetical patients that specified symptoms that are present and absent and asked participants to judge the likelihood that each patient had a target disease. Kim found a pattern of results similar to the one reported in Experiment 2. Thus, we speculate that the current results would generalize to laypersons' real-life diagnostic reasoning processes.

In the current study, we focused on laypeople's "internal" causal theories, that is, their causal theories about how the symptoms of a disorder fit together. External theories, on the other hand, might include factors such as genetics and early learning. As was mentioned at the

beginning of this article, the literature on lay theories of mental illness demonstrates that lay people show a fairly logical pattern of thinking about the external causes of disorders, often reflecting aspects of various academic theories (Furnham & Bower, 1992). However, previous research on lay theories has not investigated how the different external causes of a disorder interact with each other in the layperson's mind (i.e., Furnham, 1995; Matschinger & Angermeyer, 1996). Instead, these studies presented people with lists of possible external causes of a disorder and asked them to rate each on whether it was an actual cause of the disorder. A clear next step in research on lay theories, then, would be to examine whether laypeople's disorder categorization decisions are affected in the same way by their causal theories at a more general level. That is, would the causal status hypothesis hold up when factors outside the *DSM-IV* (APA, 1994) were included with the symptoms, such as deeper causes of the disorders as a whole? If so, the lay theories that have been reported in the previous literature as simple lists of possible causes with ratings of agreement and disagreement are much flatter and simpler than real-life lay theories.

A related question raised by the current study is also worth investigating. It is unclear whether the internal causal theories measured in the current study would remain unchanged when adding external factors. In the current studies, we did not include external factors because we focused only on the symptoms mentioned in the *DSM-IV* (APA, 1994). Intuitively, we see no reason to suspect that internal causal theories will change with an addition of external factors. For instance, it seems unlikely that a layperson would cease to report that "fear of being fat" causes "refusal to maintain body weight" because that layperson also thinks that "reading beauty magazines," for instance, causes the fear of being fat. This question awaits further research.

Generalization to experts' diagnostic reasoning

Will the results generalize to experts' diagnostic reasoning processes? Currently we are conducting a large-scale experiment using the method used in Experiment 1. We speculate that the same pattern of results will be obtained from experts. In his review of research on the cognitive processes involved in clinical inference and diagnosis, Elstein (1988) pointed out that clinicians do not, in general, make diagnoses by matching all the symptoms of the client to the lists of criteria in the *DSM-IV*. Instead, clinicians are much more likely to devote most of their attention to several symptoms thought to be "prototypical" of a specific disorder. Cantor et al. (1980) also argued that clinicians' reasoning adheres most closely to this prototype view of categorization. Our spec-

ulation is that the extra weight given to these prototypical symptoms results in part from their place in the causal structure of clinicians' theories (see Ahn et al., 2000, for the effect of features' causal status on typicality judgments). Such a link between causal theories and diagnosis in clinicians seems likely. Indeed, Mumma (1993) suggested that in diagnosis, the use of clinical theoretical knowledge of the domain precedes the use of similarity to stored representations. Likewise, Einhorn (1988, p. 53) suggested that "cues-to-causality," or cues that indicate probable causal relationships, may be used by clinicians as constraints to hypothesis construction. That is, the clinician may use these cues to narrow down possible causal scenarios for the yet-undiagnosed disorder, which at the same time narrows down the set of possible diagnoses.

Yet one might argue that the current results might not generalize to real-life clinical reasoning because we used observable symptoms, whereas real-life clinical reasoning involves many features that are inferred from observations. In line with the medical model of disorder classification, the *DSM-IV* (APA, 1994) focuses on observable symptoms. Our current study also used only observable symptoms to examine the implications of our results for the *DSM-IV* criteria. However, it is likely that experts' real-life diagnostic reasoning extends beyond the *DSM-IV* (APA, 1994). Characteristics of a disorder that are not manifested in the directly observable behavior of the patient, such as whether family members have had a similar disorder or whether a traumatic event has occurred in the patient's life, are also likely to affect the clinician's diagnostic process. Thus, it is likely that experts' real-life diagnosis incorporates a causal structure that includes features beyond those indicated in the *DSM-IV* (APA, 1994) criteria. Similarly, we categorize everyday items (i.e., apples) by their easily perceived characteristics, inferring deeper traits from these (e.g., whether an object has apple DNA; Ahn et al., 2000; Gelman & Medin, 1993; Murphy & Medin, 1985). However, we speculate that even if these nonobservable symptoms are considered in diagnosis, the causal status of symptoms would predict the importance of these nonobservable symptoms. Experiments to determine more direct answers to these questions are under way.

Can conceptual centrality determine causal centrality?

We have presented evidence that causal centrality not only predicts conceptual centrality (Experiment 1) but also influences conceptual centrality (Experiment 2). Still, there remains the possibility that conceptual centrality also affects causal centrality. Thus, a follow-up experiment might address this issue by manipulating only conceptual central-

ity. However, even if conceptual centrality appears to affect causal judgments, this would be consistent with the causal status hypothesis (Ahn et al., 2000). In our previous work, we proposed that the causal status effect results from psychological essentialism (Medin & Ortony, 1989), a construct that states that people believe concepts have essences that cause surface features (see Ahn et al., 2000, and Ahn & Kim, 2000, for more details). Thus, upon learning that a certain feature is highly central in a concept, one might also assume that that feature would be responsible for or cause other features of the concept.

Conclusions and future directions

Overall, the results of the current investigation suggest that lay categorization decisions concerning mental illness can be affected significantly by the categorizer's naive causal theories. Moreover, such an effect appears to occur in a predictable manner. We have suggested ramifications of this research for theories of categorization and lay theory research. Moreover, because a taxonomy of mental disorders is culturally determined to a great extent, the opinions of laypeople do matter for its formation. Thus, that this effect occurs in lay human reasoning could be an important research consideration for clinical theories and clinical classification into categories of mental disorders.

As a final point, we suggest several possible extensions of this research concerning relevant real-life domains. First, mental illness diagnosis is one form of person categorization, and we suggest that theory-based categorization strategies may also be used in the general domain of social categories. For instance, what we believe to be causally central features of people's personalities may determine which stereotypes concerning that person are activated. Second, because the mental illness taxonomy was influenced by the medical model, that domain may also be of interest in future investigation. Categorization decisions in medical domains ranging from neurology to dermatology may be affected by doctors' causal theories. Finally, all types of diagnostic reasoning, from diagnosis of car problems to diagnosis of computer malfunctions, may be analogous to the one studied here. That is, the effect of causal theories on diagnosis may be widely domain-general. Further research may determine whether this is the case.

Appendix A. Manipulation check for Experiment 2

To verify that the noncausal relationships with common grouping factors were interpreted as truly noncausal, an independent group of 24 participants (all undergraduate students at Yale University) received the scenarios used for

the three conditions and rated the causal strengths of the relationships between symptoms. For this task, the first-cause and last-cause information paragraphs, each of which had three causally connected symptoms and three noncausally connected symptoms, were divided so that participants considered each group of three separately. Participants considered the six symptoms of each no-cause background information paragraph together. The participants were asked, "How strong are the causal relationships between the three [or six, in the no-cause condition] symptoms (that is, one symptom causing another symptom) in the following passage?" A scale of 0–7, where 0 = *no causal relationship*, 1 = *weak causal relationship*, 4 = *medium causal relationship*, and 7 = *strong causal relationship*, was provided for each scenario.

Analysis of the pilot task data confirmed our manipulation check. A one-way ANOVA was carried out comparing the no-cause information paragraphs and only the causal parts of the first-cause and the last-cause information paragraphs. The results showed that causal strength ratings differed significantly between the three conditions, $F(2, 141) = 180.3$, $p < .001$. Planned comparisons revealed that participants rated the causal part of the first-cause scenarios as not differing in causal strength from the causal part of the last-cause scenarios, mean ratings of 5.5 and 5.7, respectively; $p = .5$. In contrast, comparisons of first-cause causal scenarios with no-cause scenarios (mean rating of 0.8, $p < .001$) and of last-cause causal scenarios with no-cause scenarios ($p < .001$) showed that both first- and last-cause causal scenarios were much greater in causal strength than in the no-cause scenarios. In addition, causal strength ratings for the first-cause and last-cause causal scenarios were reliably greater than the "medium causal relationship" midpoint rating of 4 ($p < .001$ in both cases). Interestingly, causal strength ratings for the no-cause scenarios were significantly greater than zero (mean rating of 0.8, $p < .001$). Therefore, participants were managing to extract a small degree of causal information even out of the no-cause scenarios. For our purposes, however, this does not pose a problem because the degree of causal strength in the first-cause and last-cause causal scenarios is much greater than that of the no-cause scenarios. A second one-way ANOVA comparing the no-cause information paragraphs and only the noncausal parts of the first-cause and the last-cause information paragraphs revealed no difference between noncausal grouping factors in the different conditions, $F(2, 141) = .054$, $p = .95$. Finally, two paired-sample t tests were carried out comparing the causal part and the noncausal part within the first-cause scenario (mean ratings of 5.5 vs. 0.9, respectively) and also within the last-cause scenarios (mean ratings of 5.7 vs. 0.8, respectively). In both cases, the causal part was judged to be significantly more causal than the noncausal part, $t(47) = 14.36$, $p < .001$; $t(47) = 18.6$, $p < .001$, respectively. In sum, the results of this manipulation check show that in all conditions, participants extracted very little causal information from the noncausal grouping factor. Therefore, any differences between symptoms explained by these two types of information do not result from the presence of a grouping factor but rather from causality per se.

Appendix B. Sample task for one artificial mental disorder from Experiment 2

Diagnostic criteria for Axis I disorder: Methinismus
Three (or more) of the following symptoms have been present for at least 1 month:

Intense fear of gaining weight	Recurrent unjustified suspicions of spouse's infidelity	Fake suicide attempts	Perfectionism that interferes with task completion	→	Unreasonable scorn for authority	→ Exaggerated startle response

In the diagram above, an arrow between two symptoms indicates the direction in which one causes the other. No arrow between any two symptoms means that there is no causal relationship between them.

Modern psychiatric researchers agree on the following facts about the symptoms of Methinismus:

Intense fear of gaining weight in these patients is not thought to cause or be caused by any other symptoms. Recurrent unjustified suspicions of spouse's infidelity in these patients are not thought to cause or be caused by any other symptoms. Fake suicide attempts in these patients are not thought to cause or be caused by any other symptoms. Perfectionism interfering with task completion causes these patients to unreasonably scorn authority; they resent bosses who want them to finish up. Unreasonably scorning authority, in turn, causes an exaggerated startle response in these patients because they fear being fired for their insolence and thus are always on edge.

Now, imagine to yourself a patient who has all the symptoms of Methinismus. When you have done so, please answer each of the following questions on a scale of 0 to 100, where 0 = *definitely no* and 100 = *definitely yes.*

If a patient is in all ways like a typical person with Methinismus EXCEPT that he or she does NOT have the symptom of perfectionism that interferes with task completion, does the patient have Methinismus? Your answer (0–100): _____

If a patient is in all ways like a typical person with Methinismus EXCEPT that he or she does NOT have the symptom of exaggerated startle response, does the patient have Methinismus? Your answer (0–100): _____

If a patient is in all ways like a typical person with Methinismus EXCEPT that he or she does NOT have the symptom of an intense fear of gaining weight, does the patient have Methinismus? Your answer (0–100): _____

If a patient is in all ways like a typical person with Methinismus EXCEPT that he or she does NOT have the symptom of unreasonably scorning authority, does the patient have Methinismus? Your answer (0–100): _____

If a patient is in all ways like a typical person with Methinismus EXCEPT that he or she does NOT have the symptom of recurrent unjustified suspicions of the spouse's infidelity, does the patient have Methinismus? Your answer (0–100): _____

If a patient is in all ways like a typical person with Methinismus EXCEPT that he or she does NOT have the symptom of making fake suicide attempts, does the patient have Methinismus? Your answer (0–100): _____

Notes

Nancy Kim is now at Department of Psychology, McMaster University. Woo-kyoung Ahn is now at Department of Psychology, Vanderbilt University.

Parts of Experiment 1 were reported at the 1999 Eastern Psychological Association meeting, Providence, Rhode Island. Experiments 1 and 2 were presented at the 1999 Psychonomic Society meeting, Los Angeles, California.

This project was supported in part by a National Science Foundation Graduate Research Fellowship to Nancy Kim and a National Institute of Health Grant (NIH R01-MH57737) to Woo-kyoung Ahn.

We thank Frank Keil, Peter Salovey, Steven Sloman, and Heidi Wenk for their helpful comments on this research and Andrew Tomarken for helping us use the consensus measures reported in Experiment 1. We also thank Martin Dennis, Yani Indrajana, Rebecca Shaffer, and Heidi Wenk for their help in collecting data.

Correspondence about this article should be addressed to Woo-kyoung Ahn, Department of Psychology, Vanderbilt University, 534 Wilson Hall, Nashville, TN 37240 (E-mail: woo-kyoung.ahn@vanderbilt.edu) or Nancy Kim, Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L85 4K1 (E-mail: nskim@post.harvard.edu). Received for publication February 8, 2000; revision received July 19, 2000.

1. In the actual simulation of the model, d_{ij} is one cell of a matrix D that represents all pairwise dependencies between all features of the concept.

2. We measured the conceptual centrality of the 30 criterial symptoms only and not for the characteristic symptoms. This was because, as explained earlier, the primary goal of the current study is to test our hypothesis against the prototype-like *DSM-IV* (APA, 1994), which generally does not distinguish between criterial symptoms (except for a few criterial symptoms such as depressed mood for a major depressive episode). In addition, this kept the task at a reasonable length without introducing fatigue effects.

3. Note that the averaged causal strengths reported in Figure 1 are different from the Sloman et al. (1998) model-predicted conceptual centrality scores, derived from the causal strength data. This will be described in detail later.

4. The results change only negligibly when different initial values are used.

5. Following Sloman et al. (1998), this mean correlation across categories was determined by taking Fisher's *z*-transformation of each correlation, averaging across the *z*-transform scores, and converting the mean score back to units of correlation. Because this is a composite r based on different numbers of pairs of scores, no p value can be reported.

6. The Spearman rank-ordered correlation may be lower for narcissistic personality disorder because most of the criterial symptoms are equally causally peripheral. Thus, the rank-ordering of symptoms probably was a too-conserva-

tive representation of the data. When a Pearson correlation is used, $r(9) = .70$, $p < .04$. Furthermore, our population was least familiar with narcissistic personality. In a separate pilot study reported in Kim and Ahn (in preparation), 22 Yale undergraduates rated their familiarity with about 40 *DSM-IV* (APA, 1994) disorders on seven different scales. Anorexia, depression, obsessive–compulsive personality disorder and narcissistic personality ranked 1, 2, 10, and 23, respectively, from most familiar to least familiar. The rank-ordered correlation for anorexia nervosa, though high, is not significant because there were only four criterial symptoms. Again, however, all correlations were clearly positive.

7. Gentner's account holds for relationships in general, not just causal relationships, but is still useful for the purposes of the present article because it includes causal relationships.

8. See Gentner (1989) and Holyoak (1985) for further discussion of the distinction between causal relationships and other types of relationships. We would like to thank Frank Keil for suggesting this manipulation.

9. Arguably, some of these symptoms might contain causal theories within themselves, and it seems to be human nature to draw causal interpretations from correlations. A pilot study was conducted to measure how participants perceived these materials (see Appendix A). Although control passages were not perceived to be absolutely noncausal, participants believed the relationships to be much more causal when causal information was explicitly described. Thus, our materials make a strong contrast between causal and noncausal conditions.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition, 69,* 135–178.

Ahn, W., & Kim, N. S. (2000). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *The psychology of learning and motivation.* New York: Academic Press.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. (2000). Causal status effect as a determinant of feature centrality. *Cognitive Psychology, 41,* 361–416.

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

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research, 18,* 433–447.

Blatt, S. J., & Levy, K. N. (1998). A psychodynamic approach to the diagnosis of psychopathology. In J. W. Barron (Ed.), *Making diagnosis meaningful: Enhancing evaluation and treatment of psychological disorders.* Washington, DC: American Psychological Association.

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language & Cognitive Processes, 4,* 127–155.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 458–475.

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

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: Plenum.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments & Computers, 25,* 257–271.

Einhorn, H. (1988). Diagnosis and causality in clinical and statistical prediction. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp. 51–70). New York: Free Press.

Elstein, A. S. (1988). Cognitive processes in clinical inference and decision making. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp. 17–50). New York: Free Press.

Furnham, A. (1995). Lay beliefs about phobia. *Journal of Clinical Psychology, 51,* 518–525.

Furnham, A., & Bower, P. (1992). A comparison of academic and lay theories of schizophrenia. *British Journal of Psychiatry, 161,* 201–210.

Furnham, A., & Hume-Wright, A. (1992). Lay theories of anorexia nervosa. *Journal of Clinical Psychology, 48,* 20–36.

Furnham, A., & Lowick, V. (1984). Lay theories of the causes of alcoholism. *British Journal of Medical Psychology, 57,* 319–332.

Gelman, S. A., & Medin, D. L. (1993). What's so essential about essentialism? A different perspective on the interaction of perception, language, and conceptual knowledge. *Cognitive Development, 8,* 157–168.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gentner, D. (1989). The mechanisms of analogical reasoning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* Cambridge, England: Cambridge University Press.

Heaven, P. C. L. (1994). The perceived causal structure of poverty: A network analysis approach. *British Journal of Social Psychology, 33,* 259–271.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19, pp. 59–87). New York: Academic Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: The MIT Press.

Kim, N. S. (1999). *The role of naïve causal theories on lay diagnoses of mental illness.* Unpublished master's thesis, Yale University, New Haven, CT.

Kim, N. S., & Ahn, W. (in preparation). *Clinical psychologists' theory-based representations of mental disorders affect their reasoning and memory.*

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 754–770.

Lunt, P. K. (1988). The perceived causal structure of examination failure. *British Journal of Social Psychology, 27,* 171–179.

Lunt, P. K. (1991). The perceived causal structure of loneliness. *Journal of Personality and Social Psychology, 61,* 26–34.

Lunt, P. K., & Livingstone, S. M. (1991a). Everyday explanations for personal debt: A network approach. *British Journal of Social Psychology, 30,* 309–323.

Lunt, P. K., & Livingstone, S. M. (1991b). Psychological, social and economic determinants of saving: Comparing recurrent and total savings. *Journal of Economic Psychology, 12,* 621–641.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Matschinger, H., & Angermeyer, M. C. (1996). Lay beliefs about the causes of mental disorders: A new methodological approach. *Social Psychiatry, 31,* 309–315.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* Cambridge, England: Cambridge University Press.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20,* 158–190.

Mumma, G. H. (1993). Categorization and rule induction in clinical diagnosis and assessment. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 29, pp. 283–326). San Diego: Academic Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual structure. *Psychological Review, 92,* 289–315.

Narikiyo, T. A., & Kameoka, V. A. (1992). Attributions of mental illness and judgments about help seeking among Japanese-American and white American students. *Journal of Counseling Psychology, 39,* 363–369.

Nosofsky, R. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189–228.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Wakefield, J. C. (1998). Meaning and melancholia: Why the *DSM-IV* cannot (entirely) ignore the patient's intentional system. In J. W. Barron (Ed.), *Making diagnosis meaningful: Enhancing evaluation and treatment of psychological disorders.* Washington, DC: American Psychological Association.

Westermeyer, J., & Wintrob, R. (1979a). "Folk" criteria for the diagnosis of mental illness in rural Laos: On being insane in sane places. *American Journal of Psychiatry, 136,* 755–761.

Westermeyer, J., & Wintrob, R. (1979b). "Folk" explanations of mental illness in rural Laos. *American Journal of Psychiatry, 136,* 901–905.