The Influence of Context on Categorization Decisions for Mental Health Disorders

Jessecae K. Marsh (jessecae.marsh@TTU.edu)
Department of Psychology, Texas Tech University, MS 42051
Lubbock, TX 79413 USA

Andres De Los Reyes (adelosreyes@psyc.umd.edu)
Department of Psychology, University of Maryland at College Park, 1147 Biology/Psychology Building
College Park, MD 20742 USA

Abstract

Mental health clinicians engage in an important form of real-world categorization as they diagnose their patients with mental disorder diagnoses. How are clinicians affected by the context within which diagnostic criteria of a patient present when making diagnostic evaluations? The classification system clinicians are instructed to use is structured around a statistical approach to assessing diagnosis and does allow for the interpretation of criterial features through influences like context. The following experiment tests whether clinicians are affected by the context within which non-diagnostic information about a patient is presented. We tested clinician's diagnostic judgments for symptoms of Conduct Disorder that were presented either in a context that should be perceived as being associated with Conduct Disorder or in a context that should not be perceived as being associated with Conduct Disorder. We found that clinicians were influenced by context, but in surprising ways. Clinicians lowered their diagnostic judgments for symptoms presented in a low associative context but did not change their estimates for high associative contexts as compared to baseline. The effect of context was also found to vary over the criterial symptoms that were presented, and this variation was associated with clinicians' idiosyncratic ratings of the criterial symptoms. These results have interesting implications for how clinicians view their patients and for how context affects categorization more generally.

Keywords: Categorization; clinical reasoning; theory-based reasoning.

Introduction

Mental health disorders provide an interesting domain within which to study real-world categorization. The *Diagnostic and Statistical Manual of Mental Disorders* (DSM) presents a guideline for mental health clinicians to use in the categorization of patients into mental disorder categories (American Psychiatric Association [APA], 2000). Most mental disorders included are described as lists of criterial symptoms, with membership in a given category being gained by displaying a certain number of the described criterial features. For example, in order to qualify for a diagnosis of the childhood disorder Conduct Disorder a patient must meet a series of required criteria (e.g., minimum amount of time showing the problematic behaviors, marked impairment in life functioning) and then possess any 3 of a possible 15 other criteria that typify clinically impairing levels of aggressive behavior and a disregard for rules or social norms. The exact symptoms or combination of symptoms generally do not figure into categorization as long as the required number of symptoms is met.

In order to collect information about the criterial features that are present in a given patient, mental health clinicians must go through a difficult process of gathering and synthesizing different types of information. Patients do not necessarily provide their diagnostic symptoms in a clear-cut and easy to parse fashion. Instead, such criterial features are embedded within a set of extraneous, biographical, non-diagnostic information. Although the DSM acknowledges that the context within which these behaviors occur should be taken into account when making diagnostic decisions (e.g., a child that frequently gets into fights at school for the sake of self-defense likely does not meet criteria for Conduct Disorder), such non-diagnostic contextual information that does not take the form of actual diagnostic criteria does not play a formal, codified role in diagnostic classification as delineated in the DSM. The question becomes, do clinicians use the context within which diagnostic criteria present to make their category decisions?

Evidence from the cognitive psychology field suggests that people's categorization decisions can be affected by the context in which the features are displayed. For example, Medin, Goldstone, and Gentner (1993) found that the context within which an ambiguous stimulus was presented affected the interpretation of its ambiguous features, and consequently its categorization. Similarly, Ahn, Novick, and Kim (2003) found that providing clinicians with an explanatory context for a person's mental disorder symptoms results in the person being categorized as more 'normal'. Likewise, the stereotype literature has shown that an ambiguous behavior will be categorized differently depending on contextual information such as the ethnic background of the person engaging in the behavior (Gawronski, Geschke, & Banse, 2003). Generally speaking, this evidence suggests that people are influenced by the context within which categorization information is presented.

The actual structure of DSM categories may further influence clinicians to rely on contextual information when evaluating patients. The statistical structure used in the

DSM to describe mental disorder categories does not closely align with what current research has suggested is the structure of laypeople's everyday categories. For example, research has shown that people are greatly affected by the knowledge they have about relations between features in making category decisions (e.g., Ahn, 1998; Murphy & Medin, 1995; Sloman, Love, & Ahn, 1998). Recent work has expanded these findings to mental health clinicians and found that clinicians will endorse causal relations between features of a mental disorder. Further, the presence of these relations can affect goodness-of-fit judgments for patients (Kim & Ahn, 2002). As such, clinicians appear to not always abide by the strict categorization proscribed in the DSM, and will instead use their own theories in the categorization process. This idiosyncratic form of categorization allows for the influence of context in diagnosing patients.

The following experiment investigates the influence of context on mental health professionals' judgments of hypothetical patients. Specifically, clinician participants read vignettes describing youths who have features of Conduct Disorder in a context that is associated or not associated with a diagnosis of Conduct Disorder. We chose Conduct Disorder for our materials because of the large number of criterial features (i.e., 15) that can describe the disorder category. As such, we can not only investigate the general influence of context in making diagnostic decisions, but also further investigate whether context differentially affects features within a category.

Methods

Materials

In order to test how mental health clinicians are affected by the context within which patient symptoms present, we first had to create contextual features that could plausibly be associated with a mental disorder diagnosis, but did not provide actual diagnostic information for the diagnosis. We created a set of hypothetical life factors that described youths displaying behaviors that would be perceived as being associated with a diagnosis of Conduct Disorder (High association condition) and youths displaying behaviors that would not be perceived as being associated with a diagnosis of Conduct Disorder (Low association condition). Each life factor was created in a pair that matched the basic information of the factor while varying whether the factor would be predicted to have a low or high association with Conduct Disorder. Figure 1 gives an example of such at item. The top box of this figure lists the stem that began the life factor description. The two lower boxes show the conclusion to this life factor that would describe a low or a high association with Conduct Disorder, respectively. In this way the Low and High association facts were equated in the basic premise and differed in the ending of the statement. The High association facts were worded so that they did not represent actual features of conduct disorder, but rather behaviors/traits that would plausibly occur in a child with a Conduct Disorder diagnosis. Matched pairs were created to describe three different types of life factors: child's family life, friends, and school environment.

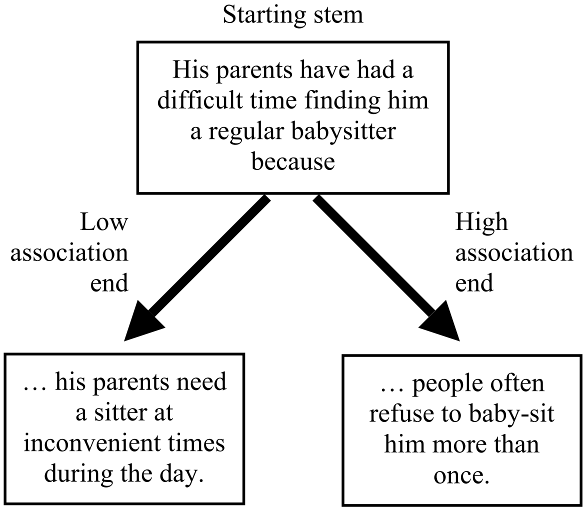
Starting stem
His parents have had a difficult time finding him a regular babysitter because
Low association end
High association end
… his parents need a sitter at inconvenient times during the day.
… people often refuse to baby-sit him more than once.

Figure 1: An example life factor.

We pretested the materials we created to a group of clinical graduate students (N=29) to ensure that the factors were associated with Conduct Disorder. We asked the pretest participants to "rate the likelihood that each fact would be present in a child with Conduct Disorder" on a 0 (not very likely) to 100 (very likely) scale. In addition, we asked the pretest participants to indicate any factors they believed were actual criterial, diagnostic symptoms of Conduct Disorder or any other mental disorder. Any factor that was indicated as being a criterial feature was dropped from use. As a first rough pass, we selected for use in further experiments feature pairs (as seen in Figure 1) that

Table 1: Example sets of Low and High association versions of a vignette.

Type of Factor	Low Association Version	High Association Version
Family	His parents have had a difficult time finding him a regular babysitter because his parents need a sitter at inconvenient times during the day.	His parents have had a difficult time finding him a regular babysitter because people often refuse to baby-sit him more than once.
Friends	His friends' parents tend to like him.	His friends' parents tend not to like him.
School Environment	He doesn't like some of his classmates because they try to cheat off his tests.	He doesn't like some of his classmates because they wouldn't let him cheat off their tests.

showed a significant difference in association scores between the high and low version as found through independent sample t-tests (*p*s < .05).

From the basic life factors that we selected through Pretest 1, we created vignettes for use in the main experiment. Each vignette consisted of one life factor randomly selected from each of three types of life factors: factors describing the youth's family, friends, and school environment. For each set of three life factors we created a High and a Low association vignette. Table 1 shows a Low and a High association vignette constructed from the same matched pairs of factors.

We conducted an additional round of pretesting on these newly constructed vignettes. We asked clinical interns (i.e., students who had already progressed past the initial years of graduate or professional training) (N=35) to rate for each vignette, "How likely would a child with the given life factors be found to have Conduct Disorder if a full clinical evaluation was given", using a scale of a 0 (not very likely) to 100 (very likely). From the results of Pretest 2, we selected 15 pairs of vignettes that showed a significant difference between their High and Low version through independent sample t-tests (*p*s < .01) for use in the actual experiment.

In the main experiment, the presentation of each pretested vignette was accompanied by the presentation of one of the fifteen features of Conduct Disorder as described in the DSM (APA, 2000). These features are as follows: *bullies others; initiates fights; used a weapon; cruel to people; cruel to animals; stolen while confronting a victim; forced someone into sexual activity; fire setting; destroyed others' property (other than by fire setting); broken into someone else's house, building, or car; lies or "cons" others; stolen without confronting a victim; stays out at night; run away from home overnight; and truant from school.* Therefore, in the main experiment participants were presented with narratives containing four features (3 life factors and one criterial conduct disorder feature). The fifteen criterial features were rotated through the vignettes such that participants read one High and one Low association vignette displaying each of the criterial features. The ratings of interest in the primary experiment were made from these four feature vignettes.

Procedure

The primary experiment consisted of two parts: a vignette rating phase, and an impression judgment phase. In the vignette rating phase, participants rated 30 separate vignettes. Fifteen of the vignettes described a High association context and 15 described a Low association context. For each vignette, participants rated, "How likely would a child with the given life factors be found to have Conduct Disorder if a full clinical evaluation was given", using a scale of a 0 (not very likely) to 100 (very likely). That is, participants made goodness-of-fit judgments for each vignette. Each vignette was presented as its own screen of the experiment. Participants had to rate a given vignette before they could move on to the next vignette. The order of the vignettes was randomized for each subject.

After the vignette phase, participants made two statistical and two theory-based judgments of the criterial features that had been presented alongside the life factors in the vignettes. Specifically, participants rated for each feature the statistical prevalence (i.e., the category validity: "what percentage of youths diagnosed with Conduct Disorder possess that diagnostic feature") and diagnosticity (i.e., the cue validity: "what percentage of youths who display that feature meet the criteria for having a diagnosis of Conduct Disorder"). The two theory-based judgments asked for ratings on the importance to diagnosis ("rate how important you believe that feature to be in diagnosing a youth with Conduct Disorder"), and the abnormality of that feature ("rate how abnormal you believe it is for the average, normal youth to possess each of the following criterial features of Conduct Disorder"). We also asked participants to perform the same likelihood judgments they did for the full vignette description on each of the criterial features alone (i.e., likelihood that a child with that symptom would receive a diagnosis of conduct disorder if a full evaluation was given). The order of the five judgments was randomized for each subject.

The experiment was designed as an online survey using the Qualtrics survey software. Participants completed the experiment at their own pace through their own home or office computer.

Participants

We recruited professional mental health clinicians who had been licensed for at least 5 years and had experience in the treatment of children (N=22) to participate in the main experiment. We contacted clinicians by posting advertisements through mailing lists of professional organizations that cater to clinicians specializing in the care of youth patients. Participants were entered into a drawing for a chance of winning a $50 online gift certificate as compensation. In addition, a $5 donation was made in each participant's name to a charity that focuses on child welfare issues.

Results

Overall Effects of Context

From a strictly DSM perspective, clinicians should not rate vignettes in the High and Low conditions differently. That is, since the vignettes were equated on the number of presented criterial features, then clinicians should rate all of the vignettes the same. However, if the context within which these diagnostic features appears affects categorization, then a difference should be found between the two conditions. This latter prediction was supported in our data. Figure 2 depicts the likelihood scores averaged across the 15 High context vignettes (*M*=54.8) in the far left bar and the scores averaged across the 15 Low context vignettes (*M*=29.2) in the far right bar. The likelihood ratings for the High and

Low conditions were significantly different, *F*(1,21) = 54.0, *p* < .001.

We further analyzed the High and Low conditions to determine how the likelihood evaluations of criterial features are affected by contextual factors when the features are presented in the vignettes as compared to the presentation of those features without context, as measured in the baseline likelihood measures. The middle bar of Figure 2 shows the average baseline likelihood judgment for the criterial features. A repeated measures ANOVA with judgment type (High, Low, vs. Baseline) as a factor found a significant main effect of judgment type, *F*(2, 42) = 25.3, p < .001. Specific paired t-tests were completed to investigate the source of this main effect. The baseline judgment (*M*=54.0) was significantly higher than the Low condition, *t*(21) = 6.40, *p* < .001. However, there was not a significant difference between the baseline judgment and the High condition, *p* = .86. In short, couching a criterial feature in a context that has a low association with Conduct Disorder reduces likelihood judgments compared to judgments for that feature alone; couching a criterial feature in a high association context does not create a change over baseline.

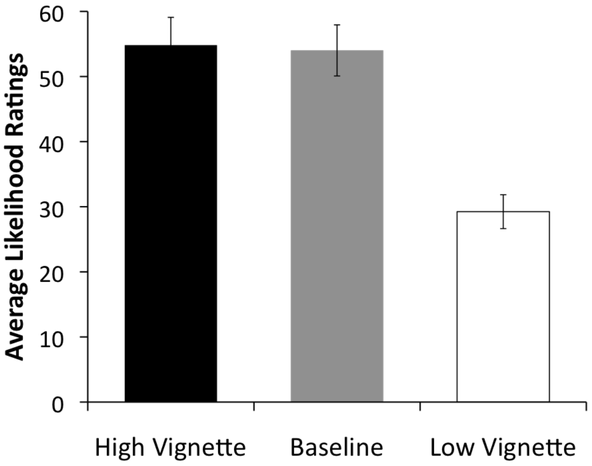

Figure 2: Average likelihood ratings for baseline compared to High and Low conditions.

Feature Specific Effects of Context

The preceding analyses collapsed ratings across all of the presented criterial features to look at the general effect of context. Using such an analysis washes out any individual differences that may exist between criterial features in their resilience to the effect of context. Looking at the criterial features listed in the Methods section, the nature of the features used as criteria for Conduct Disorder vary greatly in their severity and underlying character. For example, both *stays out at night* and *has forced someone into sexual activity* are features of Conduct Disorder that are meant to be given equal weight in diagnosis. Is it possible that features that vary so greatly in their composition and apparent severity are weighted equally by clinicians and therefore treated equally in our likelihood judgments? Our question of interest is whether these inherent differences in the nature of criterial features result in differing influences of context.

The first step in evaluating the effect of context on individual features was to create a score that could measure this influence. We calculated a context effect score for each participant's criterial feature ratings by subtracting their rating for the Low condition from their rating in the High condition for each criterial feature. A large context effect score would thus be found when context causes great shifts in ratings between the High and Low conditions, whereas small context effect scores would be found for features whose likelihood estimates do not change across conditions. To determine if these context effect scores differed across features, we rank ordered the 15 criterial features for each participant according to their context effect scores. We then averaged the context effect scores at each rank across participants. For example, we calculated the average context effect score across all participants for Rank 1, regardless of what feature it was, Rank 2, and so on. By doing this, we can compare across participants how their context effect scores varied, while ignoring what specific features were paired with high and low context presentations. For ease of comparison, we collapsed these rankings into the top five ranked scores, the middle five ranked scores, and the bottom five ranked scores. Figure 3 shows the average context effect scores for the top five (*M*=45.2), the middle five (*M*=26.4), and the bottom five (*M*=5.17) as separate bars. A repeated measures ANOVA over rank (top five, middle five, vs. bottom five) found a significant main effect, *F*(2, 42) = 102.5, *p* < .001). Paired t-tests were conducted between the three groups and significant differences (using Bonferroni corrected alpha levels) were found for all comparisons: top vs. middle: *t*(21) = 9.20, *p* < .001; top vs. bottom: *t*(21) = 11.5, *p* < .001; middle vs. bottom: *t*(21) = 7.93, *p* < .001). These differences indicate that individual criterial features of Conduct Disorder vary greatly in how they are influenced by context.

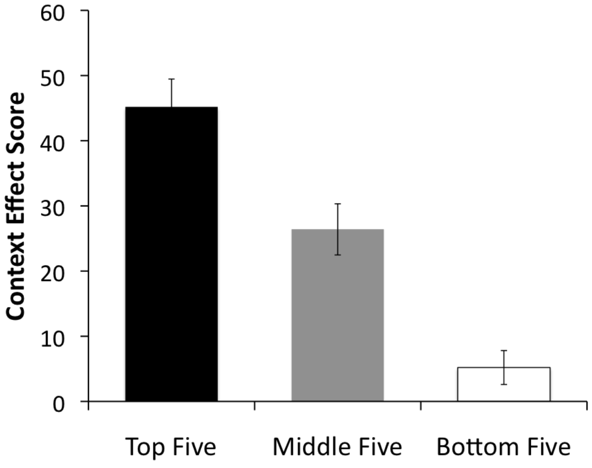

Figure 3: Context effect scores grouped by rank order.

One possible interpretation of Figure 3 is that individual features of Conduct Disorder may be influenced differently by context. That is, is there agreement among clinicians as to which features of Conduct Disorder are inherently more influenced or interpretable by the context within which they present? To answer this question, we assessed the level of agreement across participants in context scores for individual features. Participants' context effect scores were subjected to a Kendall's *W* Test to assess overall agreement between participants in ratings for individual features. Very low agreement was found, Kendall's $W = .038$, $p = .62$. This finding suggests that differences in context effect scores across features are not stable across participants. Instead, individual participants differ in what features are influenced in their judgments by context. In short, the effects of context are dependent on the clinician who is rating the feature as opposed to being a trait inherent to the criterial feature.

Our analysis so far has suggested that the effect of context differs across features in a way that is dependent on the clinician who is doing the rating. To investigate what factors may be influencing how clinicians are influenced by context for some features and not for others, we can compare each participant's context effect scores to the same participant's impression judgments (statistical and theory-based property ratings) we collected at the end of the experiment.[1] For each participant we calculated the correlation between the context effect score and the rating they provided for each criterial feature on each of the four dimensions. (Two subjects were dropped from these analyses because there was no variation in their judgments, preventing correlations from being calculated.) Averaging across participants, we found negative correlations between context effect scores and Importance ($M = -.11$) and Abnormality ($M = -.13$). These negative correlations reflect that the more abnormal or important to diagnosis a feature was rated, the less likelihood ratings differed for that feature across the High and Low contexts (i.e., the lower the context effect scores). Very small correlations were found between Prevalence (M=.04) and Diagnosticity (M=.01) and the context effect scores. To test the significance of these correlations, we used one-sample *t* tests to compare the correlations across participants to a value of zero, thereby testing if the correlations were significantly different from zero. The theory-based measures showed a relation to the context effect scores. Importance showed a correlation with context score that was significantly different from zero, $t(19) = 2.64$, $p = .016$. Abnormality approached a significant correlation, $p = .056$. The statistical measures of Prevalence and Diagnosticity were not significantly different from zero, *p*s > .5.

Discussion

The goal of the described experiment was to evaluate if clinicians are influenced by non-diagnostic context factors in assessing criterial features of mental disorder categories. In our experiment we equated the number of criterial diagnostic features and varied only the non-diagnostic features in descriptions of hypothetical youths. According to standardized diagnostic manuals used in the mental health field (APA, 2000), our vignettes should be equally likely to receive a Conduct Disorder diagnosis regardless of condition since they present only one criterial feature of Conduct Disorder and not the three needed for a diagnosis (no matter if the non-diagnostic features have a high or low perceived association with Conduct Disorder). Despite this proscription, we found that clinicians rate hypothetical youths that display non-diagnostic features that have a perceived high association with Conduct Disorder as more likely to receive a Conduct Disorder diagnosis than youths who display non-diagnostic features that have a perceived low association with Conduct Disorder. Interestingly, context appears to have an asymmetric effect across our High and Low association conditions. Specifically, the Low association condition resulted in a significant depression in likelihood ratings compared to baseline. However, the High association condition did not show any difference from baseline.

The finding that context had different effects in the High and Low association conditions is intriguing for many reasons. From a cognitive perspective, context would be expected to have symmetrical effects. That is, if a low context can reduce ratings, then a high context should increase ratings. Why did we not find such an increase in our experiment? From a mental health diagnosis point of view, one could speculate the opposite finding than what was shown in our results. Specifically, clinicians are instructed that a mental disorder diagnosis should not be provided if the diagnostic symptoms a patient displays are a normal reaction to life events (APA, 2000). With this idea in mind, the High association condition should provide a possible explanation for why the criterial feature is being displayed (see also, Ahn, Novick, & Kim, 2003). For example, perhaps having the life factors displayed in the High association condition of Table 1 would be enough reason or serve as a good explanation for why a youth would show a problematic behavior that is diagnostic for Conduct Disorder. Along this logic, the youth in the Low association condition looks particularly surprising. Why would a youth who is coming from a normal, non-problematic background as depicted in the Low association condition of Table 1 then show a behavior that is diagnostic of Conduct Disorder? With this logic, the Low association condition could in fact look more abnormal and receive a higher likelihood rating than the High association condition. It is an interesting direction for future research to determine why the High association context does not change from baseline but the Low association context does. Researching

[1] Kendall *W* was also calculated for the statistical and theory-based judgments. Prevalence (W=.514), Importance (W=.210), and Abnormality (W=.465) all showed significant agreement, all *p*s < .001. There was not significant agreement in the Diagnosticity scores, W=.062, p = .16.

this question can help illuminate how context affects categorization more generally.

In addition to the general finding of context influencing ratings overall, our results show that individual criterial features were differently affected by context. This differing effect of context for each participant was correlated with that individual's theory-based ratings for the criterial features as opposed to statistical judgments. Specifically, individual features that were impervious to the effect of context (i.e., showed the same likelihood ratings in the High and Low conditions) were given high ratings on their importance to the diagnosis process. Statistical measures, such as prevalence and diagnosticity, did not correlate with the effect of context in our ratings.

Were the measures we took the best predictors of the context effect scores? While the correlation between Importance and context effect scores was significant, it was not an overwhelmingly strong correlation. What other measures might better predict which features of a mental disorder are impervious to context and which features are not? For example, recent research has found that a clinician's emotional response to a patient can predict the therapy outcome for that patient (Marci, Ham, Moran, & Orr, 2007). Perhaps the emotional response a clinician has when reading about a problematic behavior can predict how much the context that feature is presented in will affect its interpretation. For instance, learning that a patient is cruel to animals may result in enough of a negative emotional response that the surrounding contextual features do not have the opportunity to influence diagnostic evaluations. This is a rich avenue for future research.

Summary

Mental health disorders are a unique form of category in that the structure and classification criteria for the category are explicitly laid out for the categorizer. Mental health clinicians must discover these explicit criteria from their patients amidst a large amount of non-diagnostic information that may or may not be useful for their diagnostic decision-making. Despite the explicit structure prescribed for categorization, we found that mental health clinicians are influenced by the context within which diagnostic symptoms are presented in unique and idiosyncratic ways. Future research should focus on discovering the true nature of how these individual differences in the assessment of diagnostic criteria come about. Such research will shed light on the clinical diagnostic process as well as categorization more generally.

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., Novick, L. R., & Kim, N. S. (2003). "Understanding it makes it normal:" Causal explanations influence person perception. *Psychonomic Bulletin & Review, 10*, 746-752.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders – text revision (4th ed). (DSM-IV)*. Washington, DC: American Psychiatric Association Press.

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology, 33*, 573–589

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

Marci, C. D., Ham, J., Moran, E. K., & Orr, S. P. (2007). Physiologic concordance, empathy, and social-emotional process during psychotherapy. *Journal of Nervous & Mental Disease, 195*, 103-111.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

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

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